

인공지능이 사람의 **채점**을 대신할 수 있을까?

#자동_채점 #채점_도우미 #교육_AI

곰파다 프로젝트
NLP-03조 삼각김박임



boostcamp aitech

목차

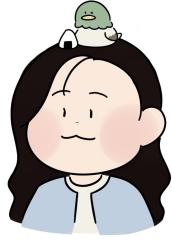
1. 소개
2. 데이터 & 모델링
3. 서비스 & 프로덕트
4. 데모
5. Summary & Discussion
6. 팀 소개

Introduction



boostcamp ai tech

안녕하세요, 삼각김박임 입니다.■

김상렬	김소연	김은기	박세연	임수정
				

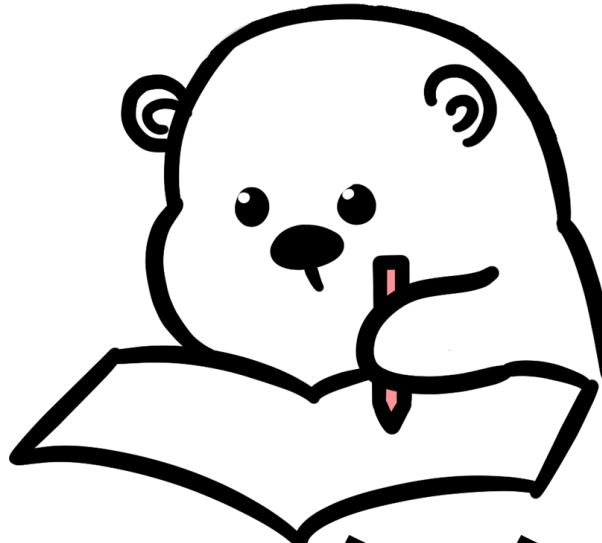
Frontend 개발
Backend API 설계
프로젝트 배포 및 관리

PM
유사도 모델링
Rule-based 데이터 구축

모델 엔지니어링
문맥 유사도 모델/
키워드 모델 제작

데이터 증강
Web/로고 디자인
전체 pipeline 연결

데이터 수집
문맥 유사도 모델
평가 데이터셋 제작
SBERT 성능 실험



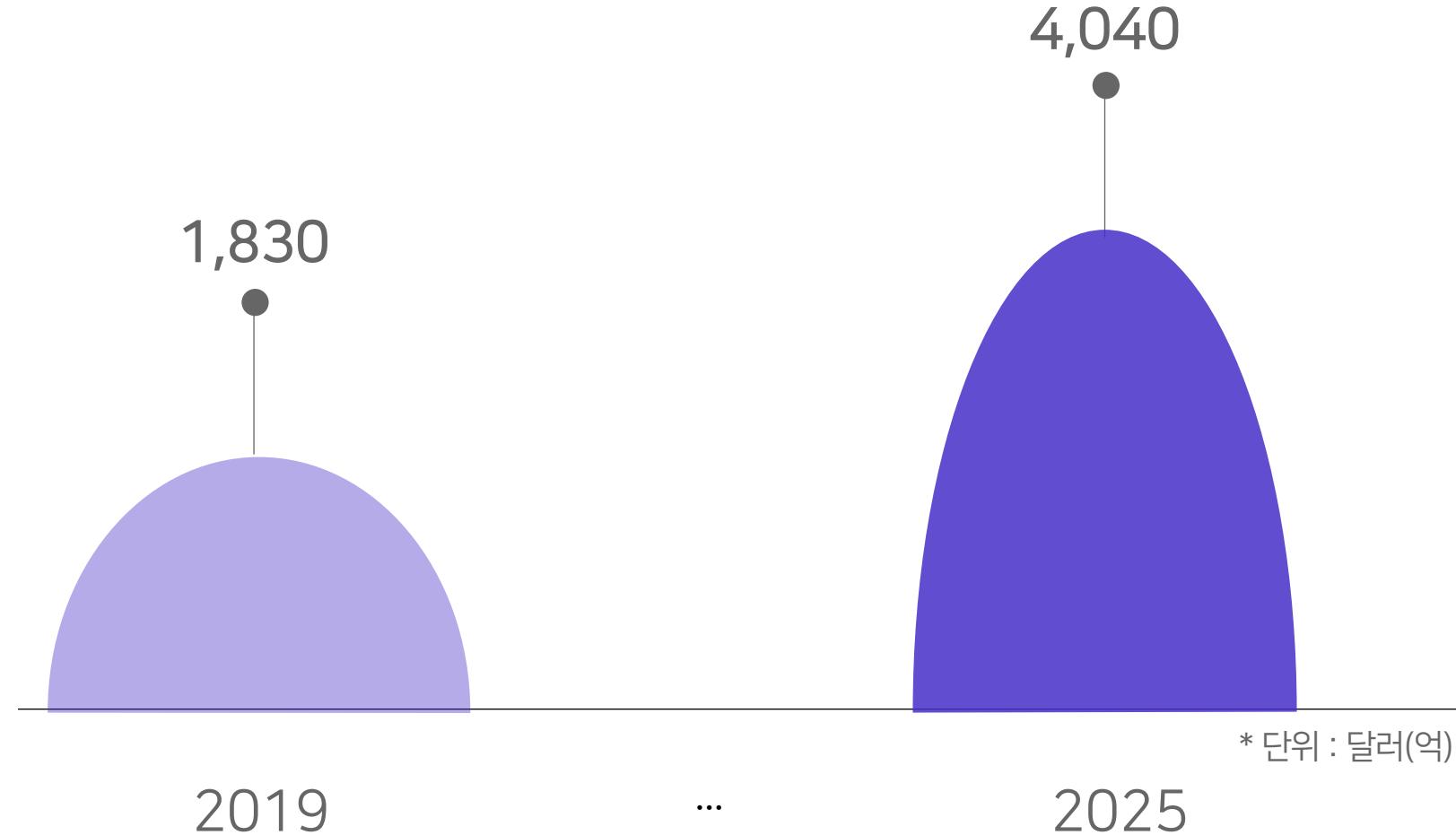
곰파다

곰파다는 채점 기준을 참고하여
서술형 답안을 꼼꼼히 채점함으로써

선생님들의 반복적 채점 작업을 효율적으로 줄여줍니다.

1. Introduction

Education + AI

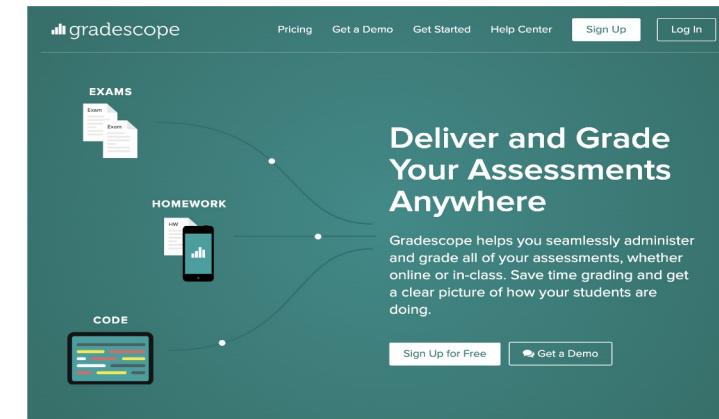


Education + AI : 해외 사례

조지아 공대에는 AI 조교가 있다[4]. 2016년 Ashok Goel 교수의 인공지능 수업에서 AI 조교 Jill Watson은 학생 질문에 답변하고, 쪽지 시험 문제를 출제하고, 토론 주제를 제시하는 역할을 수행하였다. 학생들은 이 조교가 인공지능이라고 밝혀지기 전까지 눈치 채지 못하였고, Jill Watson은 빠른 답변과 정확성으로 인기가 많은 조교였다고 한다.

정규수업이나 자기주도학습에서 학습 수준에 맞추어 수학 학습을 도와줄 수 있는 카네기 러닝 (Carnegie Learning)의 ‘MATHia’를 예를 들면, 학습자의 개념 단위 학습 과정을 파악하고 적절한 피드백과 힌트를 제공해 주는 지능형 학습지원 시스템

AI 기반 자동채점을 예를 들면, ETS에서 TOEFL iBT Speaking 시험에 2019년 8월부터 SpeechRater 기술을 적용하고 있다[7]. 사람 전문가 평가와 함께 자동평가 엔진인 SpeechRater의 점수를 합산하여 응시자에게 점수를 제공한다. 사람은 내용, 의미, 언어 전반에 대해 점수를 주고, 자동평가 엔진 SpeechRater는 발음, 억양 등의 요소를 평가한다.



Education + AI : 국내 사례

적응형 학습 서비스

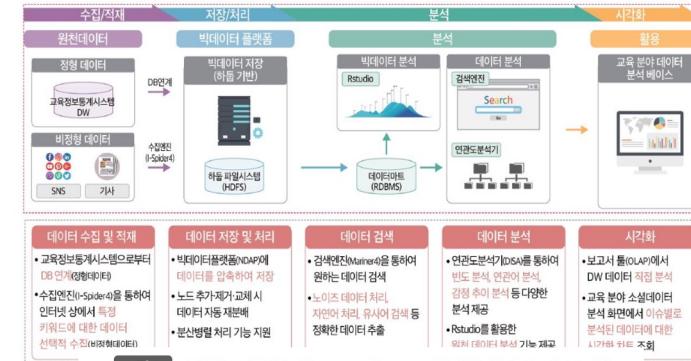
영역 구분	기업명	주요 특징
빅데이터와 인공지능을 활용한 학습 진단 및 분석	웅진 (웅진씽크빅 AI학습)	<ul style="list-style-type: none"> • 글로벌 에듀테크 기업인 Kidaptive의 ALP(Adaptive Learning Platform)를 도입하여 학습자의 흥미를 유도하면서 학습에 몰입할 수 있는 서비스 제공 • 문제 풀이시간 등 학습동행 패턴을 분석하여 학습자의 습관을 교정하는 데 도움 제공
	산타토의 (Santa TOEIC)	<ul style="list-style-type: none"> • 토익(TOEIC) 시험 점수를 빠르게 높일 수 있는 방법을 제공하는 자료 • 형 퓨터 서비스 • 빅데이터와 인공지능 기술을 활용하여 학습자의 수준을 파악하고, 이를 바탕으로 최적의 학습경로 추천 • 학습 도중 진단을 통해 학습자의 분야를 발견하면 집중적으로 콘텐츠를 제공하여 보강할 수 있도록 지원
	노리 (KnowRe)	<ul style="list-style-type: none"> • 수학 과목에 특화된 맞춤형 학습 서비스이다. 온라인 개인교사 표준으로 학습자를 대상으로 한 맞춤형 학습을 활용해 학생들의 수준에 맞는 문제를 제공하고, 반복적으로 틀리는 문제에 대해 부족한 개념을 파악하고, 이를 바탕으로 학습할 수 있는 콘텐츠 제공
	미타수학	<ul style="list-style-type: none"> • 빅데이터와 인공지능 기술을 활용하여 학습자의 취약한 수학 개념을 분석하고, 맞춤형 문제 추천 서비스 • 학습자 맞춤형으로 최적의 학습콘텐츠를 제시
	천재교육 (ilkkt)	<ul style="list-style-type: none"> • 빅데이터와 인공지능 기술을 활용하여 학습자의 학습 유형을 분석하고, 맞춤 시강포함 학습법 추천 • 유명 강사진과 강의를 제공하고, 전문가의 11학습관리 서비스 제공

자료: 한국데이터산업진흥원 「2020년 데이터산업 백서」(2020. 1)



자료: 한국교육학술정보원 「2020년 교육정보화 백서」(2020. 12)

교육 행정 지원 서비스



자료: 한



Education + AI : 국내 사례

적용형 학습 서비스

영역 구분	기업명	주요 특징
비데이터와 인공지능을 활용한 학습 진단 및 분석	(용진씽크빅 AI학습)	<ul style="list-style-type: none"> 글로벌 에듀테크 기업인 Kidaptive의 ALP(Adaptive Learning Platform)를 도입하여 학습자의 흥미를 유도하면서 학습에 몰입할 수 있는 서비스 제공 문제 풀이시간 등 학습활동 패턴을 분석하여 학습자의 습관을 교정하는 데 도움 제공
	산타토익 (Santa TOEIC)	<ul style="list-style-type: none"> 토익(TOEIC) 시험 점수를 빠르게 높일 수 있는 방법을 제공하는 지능형 투티 서비스 빅데이터와 인공지능 기술을 활용하여 학습자의 수준을 파악하고, 이에 따라 최적의 학습경로 추천
	노리 (KnowRe)	<ul style="list-style-type: none"> 수학 과정에 특화된 맞춤형 학습 서비스 빅데이터와 인공지능 기술을 활용해 학생들을 소홀히 하는 학습자에게 집중적으로 학습할 수 있도록 지원
	마티수학	<ul style="list-style-type: none"> 빅데이터와 인공지능 기술을 활용하여 학습자의 취약한 수학 개념을 분석하고, 맞춤형 문제 추천 서비스
	천재교육 (밀크)	<ul style="list-style-type: none"> 빅데이터와 인공지능 기술을 활용하여 학습자의 학습 유형을 분석하여 맞춤 시각화와 학습법 추천 유명 강사진의 강의를 제공하고, 전문가의 1:1 학습관리 서비스 제공

자료: 한국데이터산업진흥원 「2020년 데이터산업 백서」(2020. 12)

교육 행정 지원 서비스



그림 9 특성화하고 취업을 예측 프로그램 프로토 타입 설계 화면(2)-시뮬레이션

모의계산



자료: 한국교육학술정보원 「2020년 교육정보화 백서」(2020. 12)



수능에도 도입되는 서술형

수학능력평가

논·서술형 수능, 정말 도입되나?… "우려 크지만 도입 가능성 커"

尹현성 기자 | 승인 2021.04.24 07:00

| 우연철 입시전략연구소장 "이미 많은 준비됐을 것…국영수 아닌 '선택과목'만 서술형 볼듯"

[뉴스웍스=윤현성 기자] 정부가 지난 20일 발표한 '2022 개정 교육과정 추진 계획'에는 고교학점제, 원격수업 지속 활용 등이 포함됐지만 가장 주목을 끈 것은 단연 대학수학능력시험(수능) '논·서술형 시험' 도입 여부였다.

당초 이번 개정 교육과정의 골자는 2025년부터 모든 고등학교에 전면 도입되는 고교학점제인데, 이와 관련해 교육부는 2028학년도부터 적용되는 '미래형 대입제도'의 구체적인 내용을 2024년 상반기에 공개할 방침이라고 밝혔다.

수능에 논·서술형 문제가 도입될 경우 가장 우려가 큰 것은 채점 방식, 채점 소요 시간, 이의 제기 급증, 서술형 답안 채점 기준 및 공정성 문제 등이다.

수능에도 도입되는 서술형

수학능력평가

논·서술형 수능, 정말 도입되나?… "우려 크지만 도입 가능성 커"

尹현성 기자 | 2021.04.24 07:00

차점은.. AI가 하겠지?

[뉴스웍스=윤현성 기자] 정부가 지난 20일 발표한 '2022 개정 교육과정 추진 계획'에는 고교학점제, 원격수업 지속 활용 등이 포함됐지만 가장 주목을 끈 것은 단연 대학수학능력시험(수능) '논·서술형 시험' 도입 여부였다.

당초 이번 개정 교육과정의 골자는 2025년부터 모든 고등학교에 전면 도입되는 고교학점제인데, 이와 관련해 교육부는 2028학년도부터 적용되는 '미래형 대입제도'의 구체적인 내용을 2024년 상반기에 공개할 방침이라고 밝혔다.

수능에 논·서술형 문제가 도입될 경우 가장 우려가 큰 것은 채점 방식, 채점 소요 시간, 이의 제기 급증, 서술형 답안 채점 기준 및 공정성 문제 등이다.

정말, AI는 잘 채점하고 있나?

교육과정평가연구
The Journal of Curriculum and Evaluation
2014, Vol. 17, No. 2, pp. 99~122

한국어 서답형 문항 자동채점 결과 비교 분석¹⁾ - 국가수준 학업성취도 평가 국어, 사회, 과학 문항을 중심으로 -

요컨대 국어, 사회, 과학의 교과별 사용 용이 및 용례, 문항 출제 형식은 서로 다른 특징을 보이며, 이는 채점 결과에도 영향을 미쳤다. 이를 볼 때, 교과별로 지식베이스를 구축하고 이와 연계되어 차별화된 자연언어처리 및 개념 분석 기술이 정교화된다면, 현재의 단어·구 수준 자동채점 프로그램의 채점 정확성 및 효율성을 상당한 정도로 높일 수 있을 것으로 기대된다.

주제어: 자동채점, 한국어 자동채점 프로그램, 서답형 문항, 국가수준 학업성취도 평가

수능 서술형 문항 도입에 따른 인프라는 확보된 상태다. 2017년 한국교육과정평가원은 ‘한국어 서답형(주관식) 문항 자동채점 프로그램’을 개발해 특허를 취득했다. 평가원 관계자는 2016년 국가수준 학업성취도평가 표집 채점에 적용한 결과, 정확도가 100%에 가깝고 프로그램을 활용하지 않고 직접 채점했을 때보다 높은 수준이라고 밝혔다.

대학에서 자동 채점 프로그램을 주긴 하지만, 서술형에 대해서는 대부분 0점이 나온다.

모범답안과 거의 완전히 일치하는 경우에만 점수가 나온다.

띄어쓰기, 유사어, 풀어 쓴 개념에 대해 채점하지 못해 풀어 쓴 단어를 모두 제시하는 형식으로 채점하고 있다.

- 현직자 인터뷰 中

곰파다의 목표



- **다루는 과목 :** 개념어가 분명하게 드러나는 사회/과학 문제
- **채점 대상 :** 1~2문장 내외의 문장
- **채점 기준 :** 키워드 매칭, 문맥적 의미 유사도
- **최종 개발 목표 :**
 - 키워드 기준, 문맥적 의미 기준에 대한 학생 점수 제공
 - 학생들 답안에서 키워드 위치 표시

Data & Modeling



boostcamp ai tech

서술형 답안 채점 기준

내용 영역	단원	성취기준
생명	동물	혈액의 구성성분과 각 성분의 기능을 말할 수 있다.

한국 교육과정 평가원 서술형 평가 기준표 [1]

키워드 : 혈장, 적혈구, 백혈구, 혈소판

키워드 설명 :

혈장 - 물질이나 가스를 운반한다.

삼투압을 조절한다.

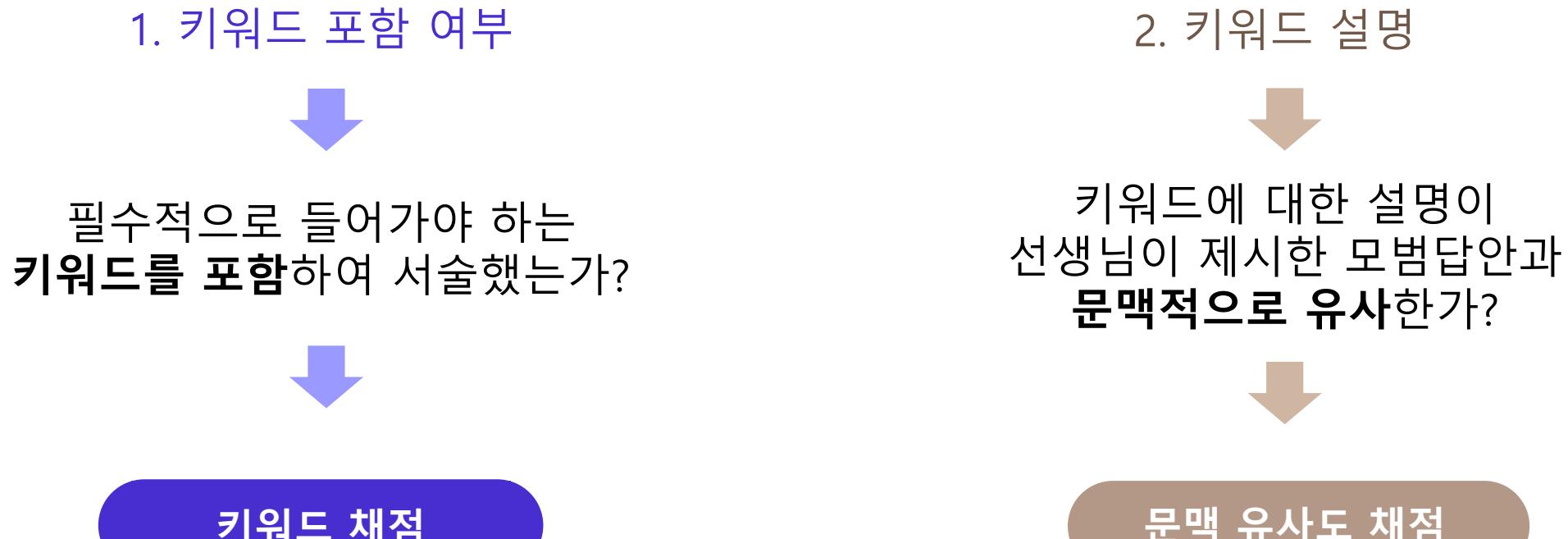
적혈구 - 산소를 운반한다.

백혈구 - 세균을 잡아먹는다.

혈소판 - 혈액을 응고시킨다.

서술형 답안 채점 기준

곰파다의 채점 기준



데이터 : 키워드 채점 & 문맥 유사도 채점

서술형 문제 및 문제에 대한 학생 답안 수집

- 정확한 키워드 존재, 옳고 그름 기준이 명확한 **초/중학교 사회, 과학 서술형 문제**로 범위를 한정하여 수집
- 가장 보편적인 [**문제(질문) + 질문에 대한 여러 개의 답변**]으로 구성된 데이터들을 후보로 수집
- 데이터 후보

학원/학교
서술형 문제

관련 연구 데이터

네이버 지식인 Q&A

데이터 : 실제 서술형 데이터

교육부와 한국과학창의재단이 지원한 서술형 평가 지원프로그램 개발 데이터 (연구책임: 하민수 교수)

2) 여름철 밖에 둔 플라스틱 병을 냉장고에 넣었을 때 찌그러지는 이유는 무엇일까요?	q38_온도	q38_압력
산소가 차갑기 때문에	1	0
공기가 차갑기 때문이다.	1	0
모름	0	0
공기를 안 냈기 때문에	0	0
녹기때문에	0	0
플라스틱병이 약해서	0	0
갑자기 뜨거운 곳에 있다가 차가운 움져서 온도변화 때문인것같다	1	0
녹아서	0	0

데이터 : 실제 서술형 데이터

문제

2) 여름철 밖에 둔 플라스틱 병을 냉장고에 넣었을 때 찌그러지는 이유는 무엇일까요?

	q38_온도	q38_압력
산소가 차갑기 때문에	1	0
공기가 차갑기 때문이다.	1	0
모름	0	0
공기를 안 뺏기 때문에	0	0
녹기때문에	0	0
플라스틱병이 약해서	0	0
갑자기 뜨거운 곳에 있다가 차가운 워져서 온도변화 때문인것같다	1	0
녹아서	0	0

데이터 : 실제 서술형 데이터

문제

2) 여름철 밖에 둔 플라스틱 병을 냉장고에 넣었을 때 찌그러지는 이유는 무엇일까요?

학생 답안

산소가 차갑기 때문에

공기가 차갑기 때문이다.

모름

공기를 안 뺏기 때문에

녹기때문에

플라스틱병이 약해서

갑자기 뜨거운 곳에 있다가 차가운 워져서 온도변화 때문인것같다

녹아서

q38_온도	q38_압력
1	0
1	0
0	0
0	0
0	0
0	0
1	0
0	0

데이터 : 실제 서술형 데이터

문제

2) 여름철 밖에 둔 플라스틱 병을 냉장고에 넣었을 때 찌그러지는 이유는 무엇일까요?

학생 답안

산소가 차갑기 때문에

공기가 차갑기 때문이다.

모름

공기를 안 뺏기 때문에

녹기때문에

플라스틱병이 약해서

갑자기 뜨거운 곳에 있다가 차가운 워져서 온도변화 때문인것같다

녹아서

q38_온도	q38_압력
1	0
1	0
0	0
0	0
0	0
0	0
1	0
0	0

키워드 &
답안 별 키워드
존재 유무

데이터 : 실제 서술형 데이터

문제

2) 여름철 밖에 둔 플라스틱 병을 냉장고에 넣었을 때 찌그러지는 이유는 무엇일까요?

학생 답안

산소가 차갑기 때문에

공기가 차갑기 때문이다.

모름

공기를 안 뺏기 때문에

녹기때문에

플라스틱병이 약해서

갑자기 뜨거운 곳에 있다가 차가운 워져서 온도변화 때문인것같다

녹아서

q38_온도	q38_압력
1	0
1	0
0	0
0	0
0	0
0	0
1	0
0	0

키워드 &
답안 별 키워드
존재 유무

서술형 문제 개수	총 64문제
문제 별 학생 답변 수	850 ~ 1000개

데이터 : 키워드 채점

Train

- ✓ 키워드 확인 방법 후보 :
TF-IDF, BM25, Word Embedding ...
- 학생 답안 데이터를 활용해
모델에 필요한 corpus 수집

Validate & Test

서술형 문제에 대한 학생 답안 + 문제별 키워드

2) 여름철 밖에 둔 플라스틱 병을 냉장고에 넣었을 때 찌그러지는 이유는 무엇일까요?

키워드

학생 답안	
산소가 차갑기 때문에	
공기가 차갑기 때문이다.	
모름	

q38_온도	q38_압력
1	0
1	0
0	0

데이터 : 문맥 유사도 채점

Train

STS(Sentence Textual Similarity) Task

- 오픈 데이터

	KorSTS[1]	paraKQC[2]	Kor-sentence[3]	KLUE STS[4]
라벨	0~5 값	0 또는 1	0 또는 1	1. 0 또는 1 2. 0~5 값
특징	짧은 문장. 외국 STS-B 번역. 뉴스, 표현 설명 내용	짧은 문장. 질문중심	짧은 문장. 지식인 질문 포함. 인터넷 용어 다수	짧은 문장. Airbnb, Policy, paraKQC 포함
데이터 개수	5,749	15,170	61,220	11,668

- Rule based 로 유의어, 반의어 데이터 제작

- 라벨 : 0 또는 1
- 특징 : 짧은 문장 / 유의어, 반의어 Pair를 만들어 데이터 제작
- 데이터 개수 : 14,390

데이터 : 문맥 유사도 채점

Validate & Test

수집한 서술형 데이터의 문제

- 모범 답안 없음
- (모범 답안, 학생 답안) Pair 에 대한 유사도 label 없음



각 문제에 대한 **모범 답안 제작** 및
(모범답안, 학생 답안) **유사도 라벨링** 작업 진행

데이터 : 문맥 유사도 채점

Validate & Test

1. 문제 별 모범답안 제작

문제 : 분해자(버섯, 곰팡이)가 지구에서 사라지면 어떤 일이 생길까요?

키워드 : 사체, 배설물, 생태계

참고 EBS 교육 자료[1] :

(3) 생산자나 분해자가 없어진다면 생태계에 일어날 수 있는 일

② 분해자가 없어진다면 죽은 생물과 생물의 배출물이 분해되지 않아서 우리 주변이
죽은 생물과 생물의 배출물로 가득 차게 될 것입니다.

제작한 모범답안

분해자가 없어 진다면 **죽은 생물**과
배설물이 분해되지 않아서
생태계가 오염될 것이다.

데이터 : 문맥 유사도 채점

Validate & Test

2. (모범 답안, 학생 답안) Pairing & Pilot Tagging 진행

키워드 : 사체, 배설물, 생태계

제작한 모범 답안 : 분해자가 없어진다면 죽은 생물과 배설물이 분해되지 않아서 생태계가 오염될 것이다.



학생 답안	평가자 1	평가자 2	평가자 3	평가자 4	평가자 5
<u>죽은 생물</u> 이 썩지 않고 그대로 있고 <u>쓰레기</u> 도 분해되지 않는다. <u>생태계</u> 가 멸종될 것이다.	2점	2점	3점	2점	2점
	1점	0점	1점	0점	0점

* 라벨링 기준표 appendix 참고

데이터 : 문맥 유사도 채점

Validate & Test

3. 반의어를 고려한 negative sample 추가

Positive sample

제품 질이 향상되고
저렴한 가격에 상품이 팔릴 것이다.

Label : 1



Negative sample

제품 질이 낮아지고
비싼 가격에 상품이 팔릴 것이다.

Label : 0

데이터 : 문맥 유사도 채점

Validate & Test

문맥 유사도 모델 평가 데이터 총 110개 제작 완료

(제작 데이터 예시)

문제	모범 답안	학생 정답	Label
q55	분해자가 없어진다면 죽은 생물과 배설물이 분해되지 않아서 생태계가 오염될 것이다.	죽은 생물이 썩지 않고 그대로 있고 쓰레기도 분해되지 않는다.	1
q55	분해자가 없어진다면 죽은 생물과 배설물이 분해되지 않아서 생태계가 오염될 것이다.	생태계가 멸종될 것이다.	0
q51	제품의 가격이 낮아지고, 품질이 좋아진다.	제품 질이 향상되고 저렴한 가격에 상품이 팔릴 것이다.	1
q51	제품의 가격이 낮아지고, 품질이 좋아진다.	제품 질이 낮아지고 비싼 가격에 상품이 팔릴 것이다.	0

데이터 제작 추가 : 문맥 유사도 채점 Pseudo Labeling

- 데이터 제작의 효율성 문제!

사람이 직접 모범답안 제작 + 유사도 채점 → 시간 대비 완성 데이터 개수가 너무 적다.

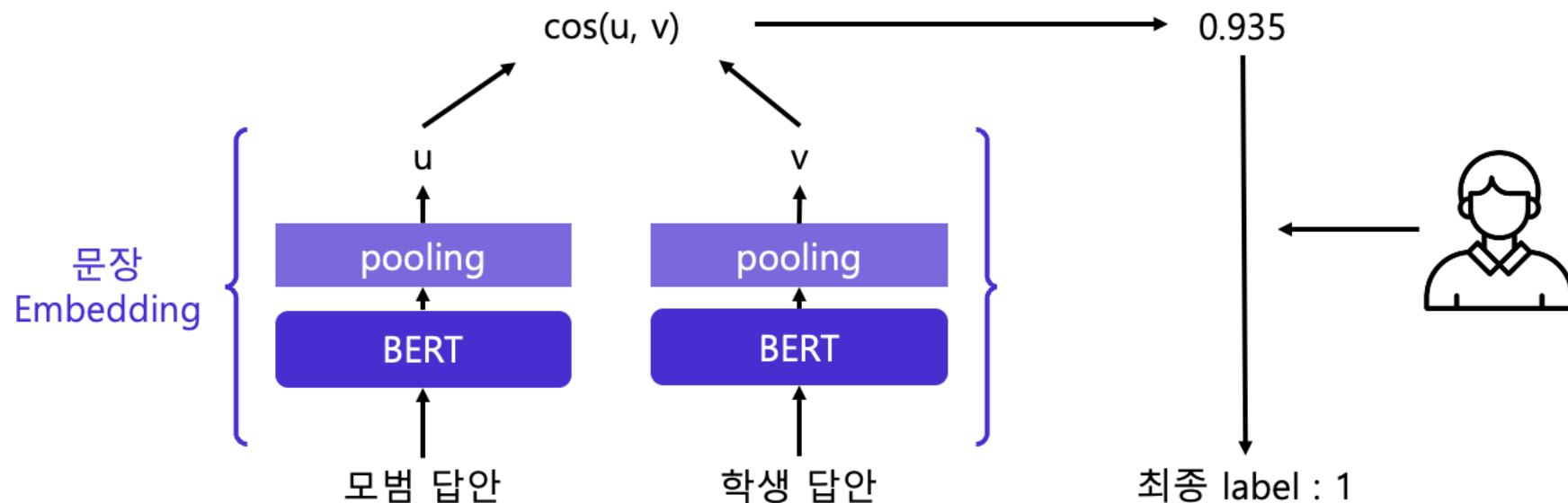
데이터 제작 추가 : 문맥 유사도 채점 Pseudo Labeling

- 데이터 제작의 효율성 문제!

사람이 직접 모범답안 제작 + 유사도 채점 → 시간 대비 완성 데이터 개수가 너무 적다.



학습된 문장 임베딩 모델을 활용해 코사인 유사도 계산 후, 사람이 최종 라벨링



데이터 제작 추가 : 문맥 유사도 채점 Pseudo Labeling

모범 답안 예시 : 나무 재질보다 금속 재질이 열 전달이 더 잘 되기 때문이다

모범 답안과 학생 답안의 코사인 유사도 기준으로 학생 답안을 정렬한 결과

<u>상위 20% 이내</u> 학생 답안 예시	<u>하위 20%</u> 학생 답안 예시
금속이 나무보다 열을 더 잘 흡수를 잘하기 때문이다. 나무보다 금속이 열 전달을 더 잘하기 때문에. 나무 국자보다 금속 국자가 열 전도성이 좋기 때문이다. 나무 국자보다 금속 국자가 더욱 열이 잘 전달되기 때문 이다.	열이 통하지 않아서 그냥 그러기 때문에 나무는 전도가 된다 모르겠습니다 안 배웠어요.

Data & Modeling



boostcamp ai tech

최종 채점 기준

질문 : 여러 제과점이 서로 경쟁을 하면 소비자에게 어떤 점이 좋을까요?

키워드 채점

키워드 :

가격, 품질, 다양성, 혜택

학생 답안 :

빵을 더 낮은 값에 살 수 있고, 빵의 맛이 더 좋아져서 소비자들은 많은 혜택을 받는다.

키워드 점수 : $3/4 = 0.75$

문맥 유사도 채점

모범 답안:

제품의 가격이 낮아지고, 품질이 좋아진다. 또한 제품의 다양성이 증가하므로 소비자들은 더 좋은 혜택을 받을 수 있다.

학생 답안 :

빵을 더 낮은 값에 살 수 있고, 빵의 맛이 더 좋아져서 소비자들은 많은 혜택을 받는다.

유사도 점수: 0.82

$$\text{최종 점수} = \text{키워드 점수} * \alpha + \text{유사도 점수} * (1 - \alpha)$$

곰파다 모델링 전체 구조

키워드 확인 모델 f

문맥 유사도 채점 모델 g

$$f(key\ words, sentence\ A) = score$$

$$g(sentence\ A, sentence\ B) = score$$

곰파다 모델링 전체 구조

키워드 확인 모델 f

문맥 유사도 채점 모델 g

$$f(key\ words, sentence\ A) = score$$

$$g(sentence\ A, sentence\ B) = score$$

키워드 채점 : 키워드 기반 분류

- HARD Exact Match 가 충분할까?
 - ① 가장 엄격한 기준은 Exact Match! BUT 유의어를 고려할 수 없음

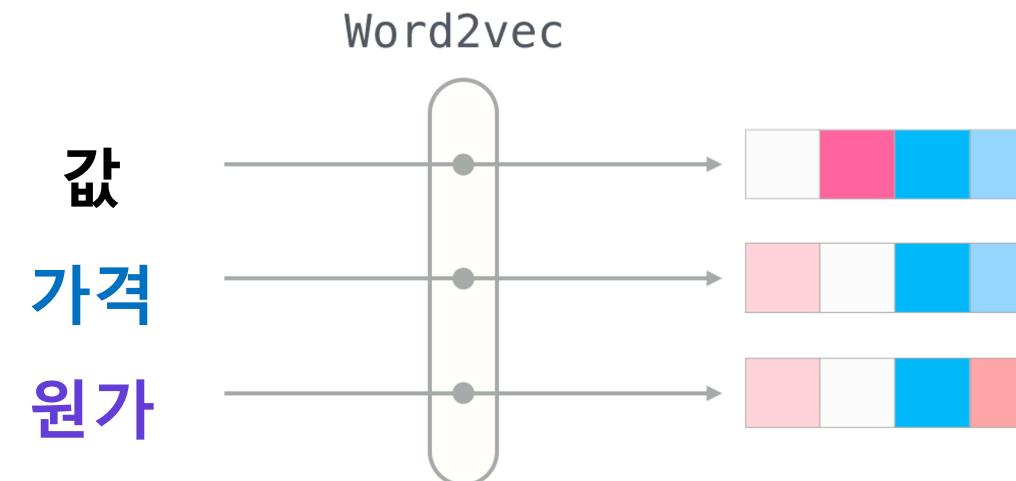
키워드 : 가격

ex1) 부덕이의 답안 : 어느 곳에 빵의 가격이 낮은 지 알 수 있다.

ex2) 메타몽의 답안 : 어느 곳에 빵의 값이 낮은 지 알 수 있다. (매칭 실패)

키워드 채점 : 키워드 기반 분류

- 단어의 임베딩!? 단어를 벡터로? **Word2Vec**?



키워드 : **가격**

들어온 문장 : 어느 곳에 빵의 값이 낮은지 알 수 있다.

가격과 유사한 단어 : **값**

키워드 채점 : 키워드 기반 분류

- Top-K보다는 일정 유사도 값 이상 넘는 단어만 정답으로 채택!

부덕이 1	모르겠다. 소비자라는 걸 몰른다 ㅇㅈ???????
부덕이 2	돈을 벌 수 있다.
부덕이 3	광고지
부덕이 4	공짜를 좋아하면 대머리됨ㅋ

키워드 : 가격

들어온 문장 : 공짜를 좋아하면 대머리됨ㅋ

Top-1 선택 시 공짜가 선택될 수 있음.

곰파다 모델링 전체 구조

키워드 확인 모델 f

문맥 유사도 채점 모델 g

$$f(key\ words, sentence\ A) = score$$

$$g(sentence\ A, sentence\ B) = score$$

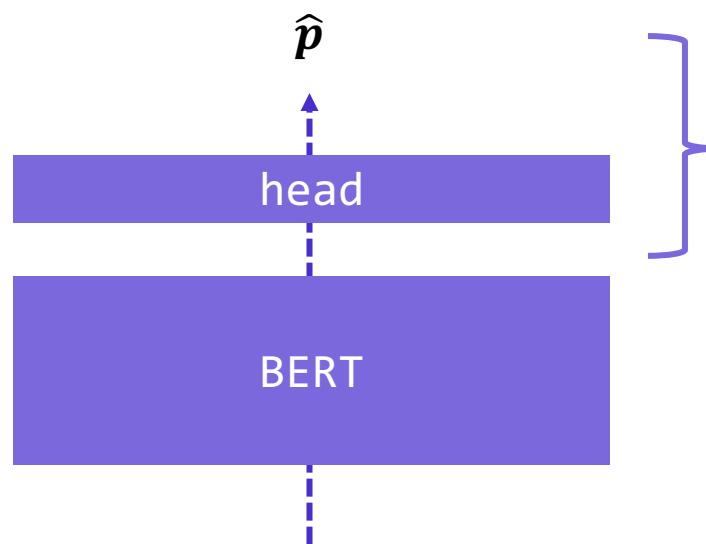
문맥 유사도 채점 : 문장 유사도 판단 학습

- 기본적인 문장 유사도 학습 방식
 - 비교할 두 문장 합쳐 모델의 입력으로 넣어주고 회귀 또는 분류 학습



문맥 유사도 채점 : 문장 유사도 판단 학습

- 유의어, 반의어와 긴 문장으로 구성된 실제 구축 validation에서 낮은 성능
- 문장 비교를 잘 하고 있는 것일까?



우리 곰파다 최고 + [SEP] + 우리 곰파다는 훌륭하네

BERT 모델	Threshold			
	출력 확률이 threshold 이상일 시 동일 문장 예측으로 처리, binary accuracy	0.6	0.7	0.8
오픈 벤치마크 validation	0.93	0.92	0.92	
실제 구축 validation	0.62 (0.31▼)	0.60 (0.32▼)	0.48 (0.44▼)	

문맥 유사도를 어떻게 효과적으로 파악할까?

- 각 문장 임베딩의 유사도를 비교하는 SentenceBERT 구조 사용
 - 추가 튜닝 없이 xlm-bert-base-pretrained로 1차 성능 비교

* xlm : cross-language model

$$Loss = (label - \text{Similarity}(\hat{p}, \hat{q}))^2$$

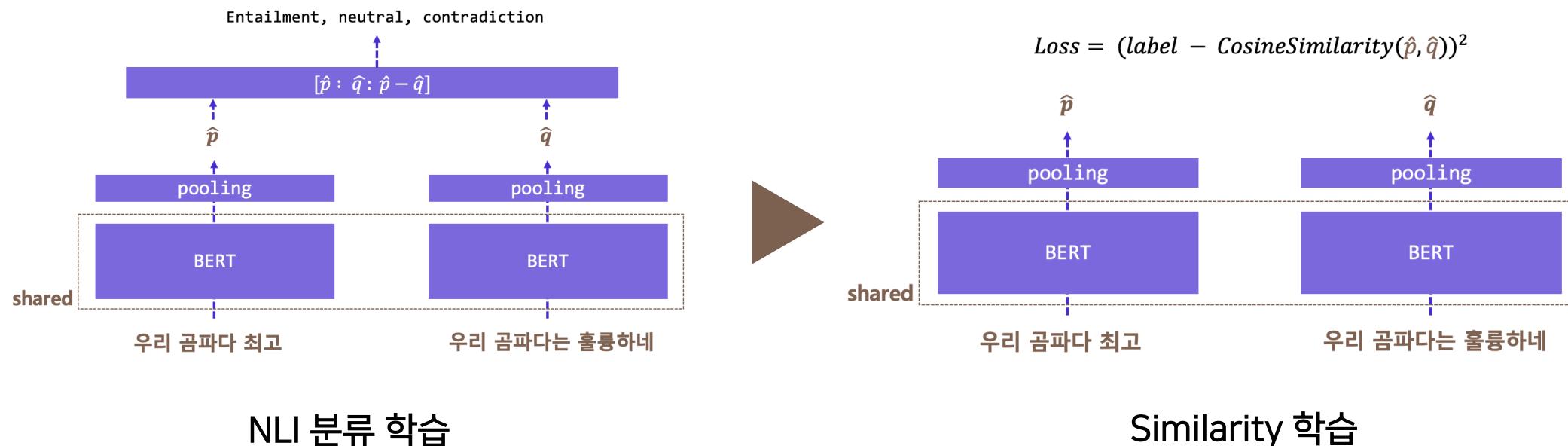


실제 구축 validation	Threshold		
	Similarity가 threshold 이상일 시 동일 문장 예측으로 처리 binary accuracy		
	0.6	0.7	0.8
BERT	0.62	0.60	0.48
SBERT	0.64 (0.02▲)	0.59 (0.01▲)	0.62 (0.14▲)

문맥 이해를 더 잘할 수 있게 해보자

① 한국어 Pretrained 모델 사용

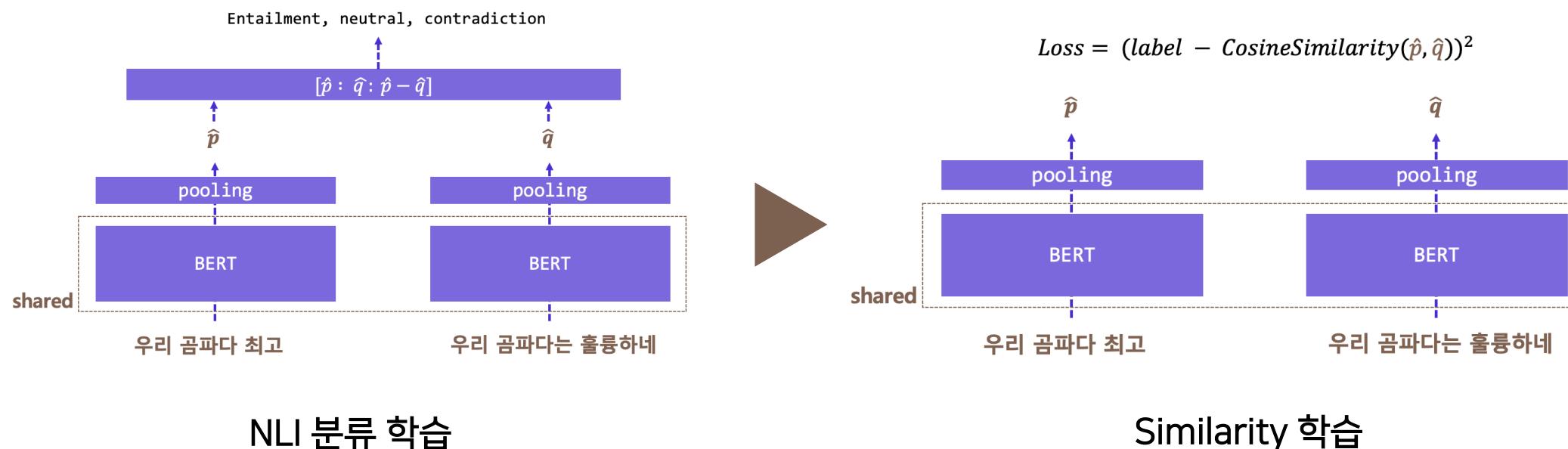
- 한국어 이해에 특화된 모델 필요 → 한국어로 사전 학습된 모델 사용



문맥 이해를 더 잘할 수 있게 해보자

② 문맥 비교에 도움될 수 있는 사전 테스크 학습

- 두 문장의 연결과계를 추론하는 Natural Language Inference(NLI)로 사전 fine-tuning 후 Semantic Text Similarity(STS) 학습



문맥 이해를 더 잘할 수 있게 해보자

- BERT 구조 대비 향상된 성능의 문맥 유사도 모델 구축 완료!
 - SBERT 구조 + 한국어 pretrained 모델 + NLI + STS

실제 구축 Validation	Threshold		
	0.6	0.7	0.8
BERT	0.62	0.60	0.48
SBERT (xlm-bert-base/ 추가 튜닝 X)	0.64	0.59	0.62
SBERT +STS (xlm-bert-base)	0.66	0.61	0.59
SBERT +NLI+STS (klue/bert-base)	0.67 (최종 : 0.05▲)	0.65 (최종 : 0.05▲)	0.58 (최종 : 0.10▲)

문맥 이해를 더 잘할 수 있게 해보자

- BERT 구조 대비 향상된 성능의 문맥 유사도 모델 구축 완료!
 - SBERT 구조 + 한국어 pretrained 모델 + NLI + STS

모범 답안	경쟁사가 있을 경우 서로의 상품이 잘 팔리게 하기 위해 가격도 낮추고 상품의 품질도 좋아진다. 또 상품의 다양성을 늘리는데도 도움이 된다.	BERT 스코어	SBERT 스코어
답변 1	제품의 가격이 낮아지고, 품질이 좋아진다. 또 제품의 다양성이 증가하고, 소비자들은 더 좋은 혜택을 받을 수 있다.	0.56	0.81
답변 2	서로의 이권을 더 얻기 위해 품질이나 가격경쟁력 따위를 높이기 위해 노력하여 소비자는 더 질 좋으면서도 값싼 상품을 얻을 수 있을 것이다.	0.52	0.75

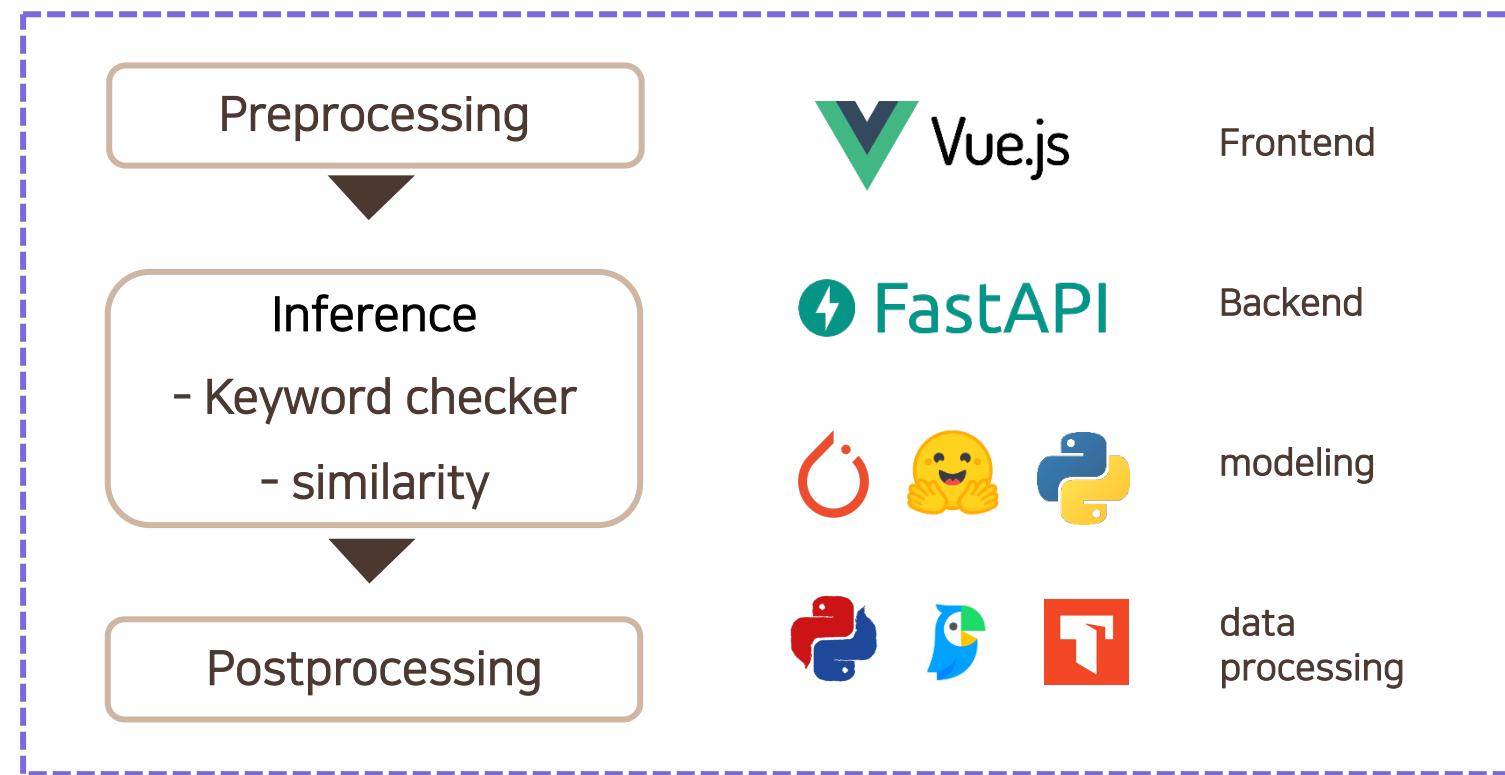
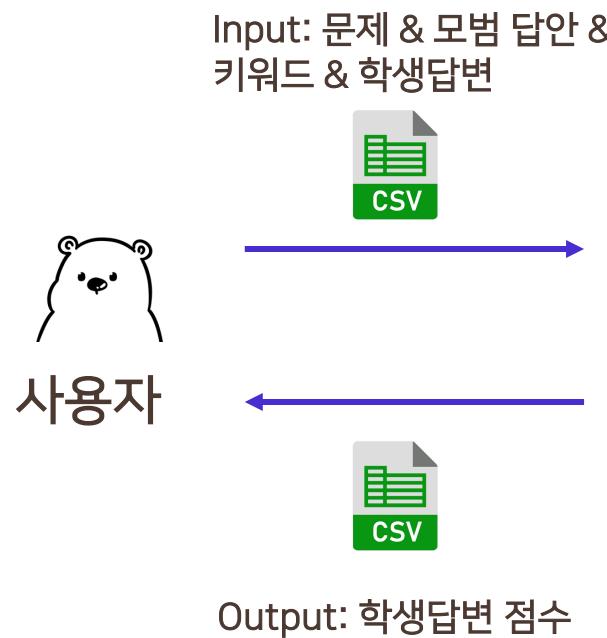
* 추가 예시 appendix 참고

Service & Product



boostcamp aitech

시스템 구조



시스템 프레임워크 선정

Frontend 와 Backend의 분리	<p>Streamlit등의 라이브러리는 사용자 커스텀 환경 부족 MVC 패턴으로 분리 어려워 협업의 난이도가 높음</p> <p>커스터마이징 및 업무 분리를 효과적으로 하기 위해 분리</p>
Vue.js	<p>중규모의 프로젝트 프로토타이핑을 위해서 도입 다양한 외부 라이브러리(NPM)과 그래프, 차트를 적용하기 위해 사용</p> <p>빠른 프로젝트 프로토타이핑, 필요한 기능 구현 위한 외부 라이브러리</p>
FastAPI	<p>가볍고 강력한 ASGI 툴킷인 Starlette 사용, API Docs를 통한 API 문서 공유 용이, 모델 서빙과 API 단의 구조적 설계를 통한 코드 관리 용이</p> <p>문서화와 구조적 설계 용이</p>

코드 최적화



```
@app.on_event("startup")
async def modelUp()
```

• 메모리 캐싱

- FastAPI 의 `app.on_event`를 통하여 model, tokenizer등을 server 시작 시에 **메모리에 올려두고 run_forever 진행**
- 모델 재로딩할 불필요, API 호출시에 inference만 진행 : **20초**에서 **0.5초**로 줄임 (* 동일 인풋 사이즈로 측정)

• 동기/비동기 처리

- `async/await` 함수 도입으로 **서버의 worker 숫자까지 비동기적으로 처리**, 동기 처리 대비 성능 향상을 기대해 볼 수 있음

Demo



boostcamp aitech

곰파다 시현

- 기본 정보 추가
- 문제 추가 디테일
- 결과 확인
- 결과 확인 상세

Summary & Discussion

boostcamp ai tech

곰파다를 요약하자면

실제 교육현장에서 선생님들의 채점 환경 반영

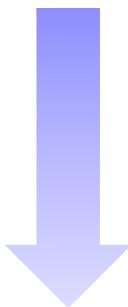
문맥 유사도
채점

키워드 채점

		Animal ID and score				
Monitoring parameter	Severity					
Body weight loss control weight = _____ (assess before fluid administration)	Control weight	0				
	<10% weight loss	1				
	10–20%	2				
	>20%	3				
Mentation/exploratory behavior	Normal	0				
	Less alert, mild depression, rough hair coat	1				
	Dull, depressed, rough hair coat	2				
	Moribund, unresponsive to handling	3				
Respiration	Normal	0				
	Mild change in rate	1				
	Consistently increased rate	2				
	Shallow or labored breathing	3				
Ambulation/paralysis	Normal	0				
	Mild lameness	1				
	Moderate impairment, sternal but ambulates with difficulty	2				
	Not moving without prompting	3				
Autophagia/dermatitis	Licking of 1 or more paws or surgical site	0				
	Open sores or bleeding	1				
	Chewing of one or more paws with no bone exposed	2				
	Chewing of one or more paws with bone exposure	3				
Dehydration (assess before fluid administration)	Normal skin turgor	0				
	Mild skin tent and decreased turgor	1				
	Moderate skin tent	2				
	Severe skin tent, sunken eyes	3				
Monitoring guidelines	Daily assessment if any score >0 Notify veterinary staff if score in any category is >1					
Euthanasia guidelines	Euthanasia recommended if any of the following conditions are met and two individuals (investigators or veterinary staff) have assessed the animal and agree on the score. Euthanasia if: 1. Score of 3 in any two categories 2. Score of 3 in any one category after day 5 postoperative 3. Score of 3 at any time for autophagia/dermatitis					

곰파다를 요약하자면

단순
Rule-Based Model



딥러닝
채점 모델

"배설물이 많아지고 동물의 사체도 많아진다"

"동식물의 시체가 분해되지못함"

"죽은 동물들의 시체와 배설물들이 섞지 않고 그대로 있을것이다"

키워드 :

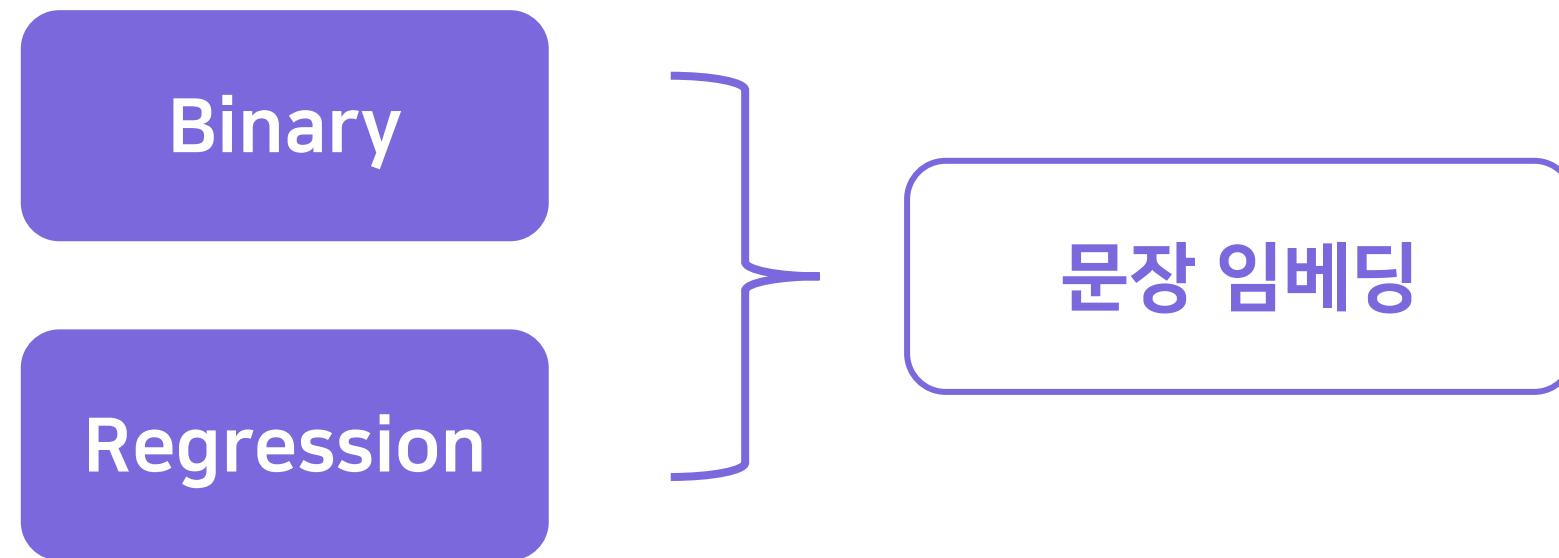
가격, 품질, 다양성, 혜택

학생 답안 :

빵을 더 낮은 값에 살 수 있고, 빵의 맛이 더 좋아져서 소비자들은 많은 혜택을 받는다.

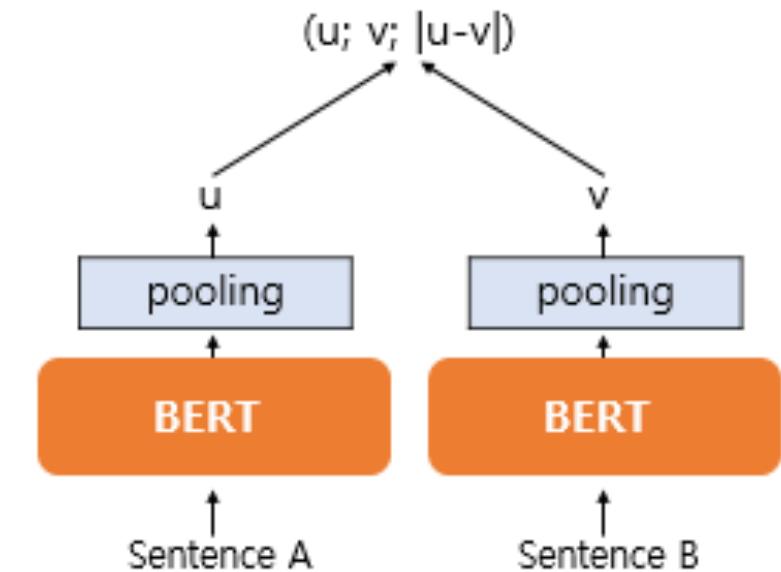
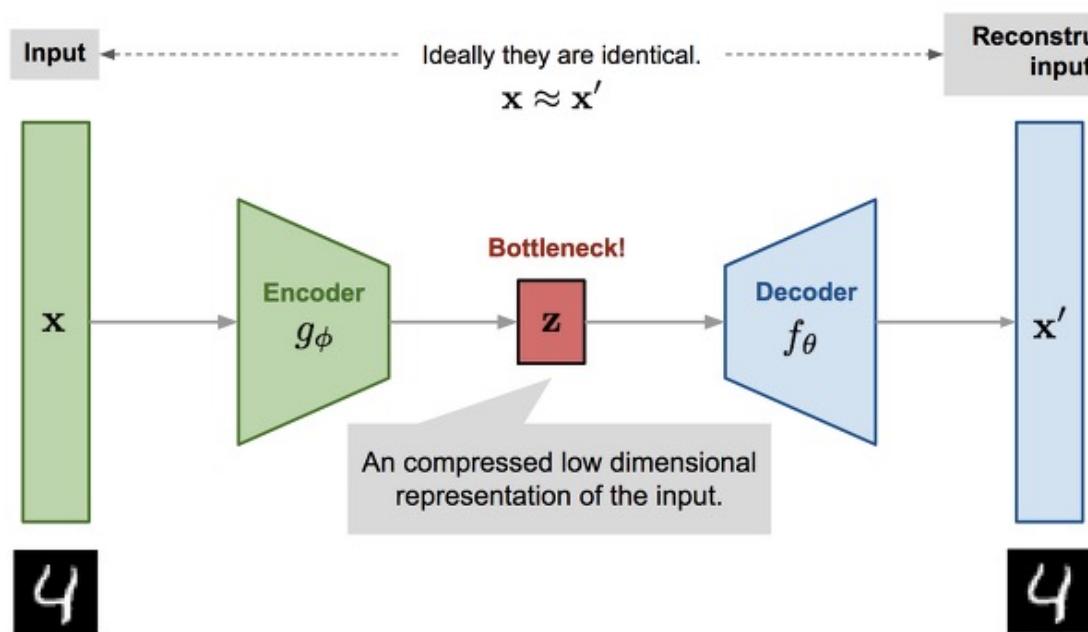
곰파다의 Further Works : 모델

문장의 Embedding을 어떻게 더 효과적으로 할 수 있을까?



곰파다의 Further Works : 모델

문장의 Embedding을 어떻게 더 효과적으로 할 수 있을까?



곰파다의 Further Works : 모델

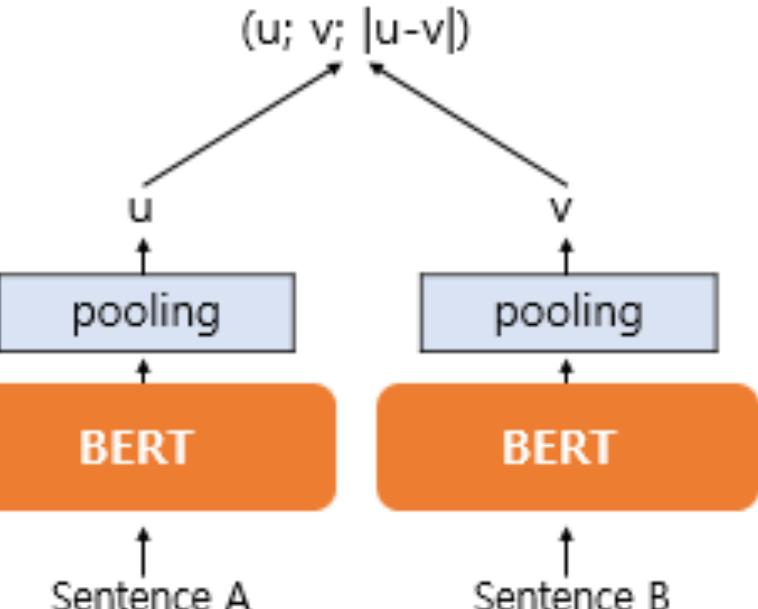
문장의 Embedding을 어떻게 더 효과적으로 할 수 있을까?

영희의 답안 : 고도가 높아짐에 따라 대기압이 낮아진다.

문장 분해

고도가 높아짐에 따라

대기압이 낮아진다.

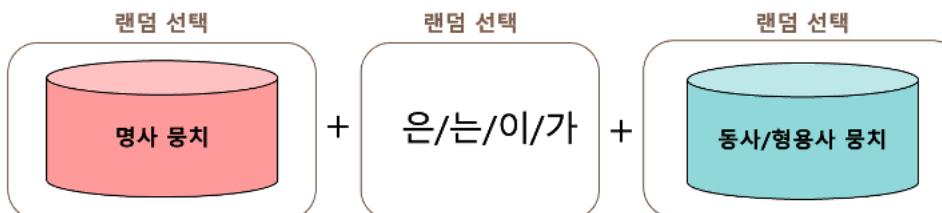


곰파다의 Further Works : 데이터

- 답안의 특성 상 반의어, 유의어를 잘 파악할 수 있는 모델 필요!
- 시도한 점 : 실제 문제 및 데이터 내 빈도 높은 명사, 동사로 의미, 유의어 및 반의어 구축, 이를 통한 rule-based 데이터 제작 시도

영희의 답안 : 고도가 높아짐에 따라 대기압이 낮아진다.

철수의 답안 : 고도가 올라가면서 기압이 높아진다.



생성된 원래 문장 : 하늘은 푸르다

Positive Samples

지평선이나 수평선 위로 보이는 무
한대의 넓은 공간은 푸르다

Negative Samples

하늘은 안 푸르다
하늘은 누렇다

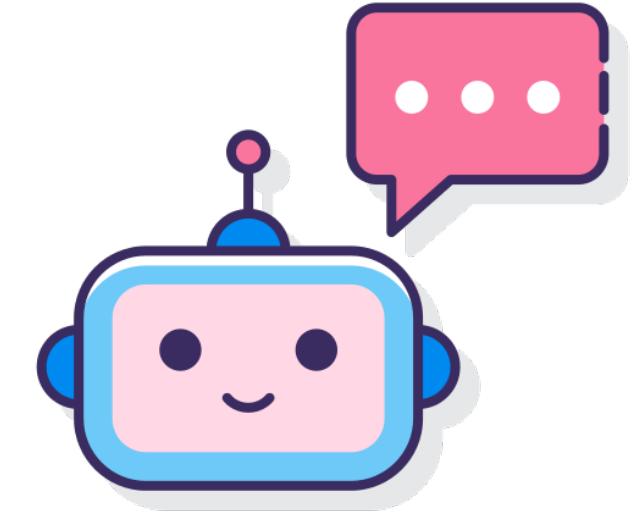
곰파다의 Further Works : 데이터

- 한국어 **반의어 사전의 부족**으로 여러 변환 단계가 필요했음
- 반의어 변환 과정에서 **품사가 변화하는 경우**가 있었다.
- 긴 문장에 따른 데이터 구축 부족, 실수 **라벨링 기준 모호**



곰파다의 Further Works : 지표, 시스템

- 채점 모델을 위한 평가, 지표의 부재
 - 키워드 모델 threshold의 휴리스틱한 설정
 - 모범답안과의 유사도를 위한 Metric 부재
- 시스템
 - 선생님들의 피드백 기반 학습을 위한 MLops 구현 부족



About US 😊



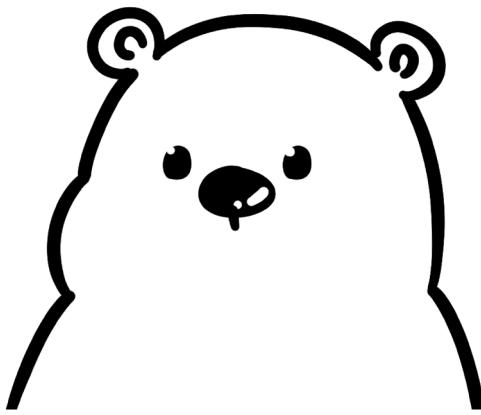
boostcamp aitech

팀 소개! + Roles

김상렬	김소연	김은기	박세연	임수정
				
Frontend 개발 Backend API 설계 프로젝트 배포 및 관리	PM 유사도 모델링 Rule-based 데이터 구축	모델 엔지니어링 문맥 유사도 모델/ 키워드 모델 제작	데이터 증강 Web/로고 디자인 전체 pipeline 연결	데이터 수집 문맥 유사도 모델 평가 데이터셋 제작 SBERT 성능 실험

감사합니다

Q&A



Appendix



boostcamp ai tech

데이터셋 추가 레퍼런스

- [1] <https://github.com/kakaobrain/KorNLUDatasets>
- [2] <https://github.com/warnikchow/paraKQC>
- [3] <https://github.com/yoongi0428/Kor-Sentence-Similarity>
- [4] <https://klue-benchmark.com/tasks/67/data/description>

Training & Deployment 환경

- Training & Deployment 서버 환경

- [OS] Ubuntu 18.04.6 LTS
- [CPU] Intel(R) Xeon(R) Gold 5220 CPU @ 2.20GHz
- [GPU] Tesla V100-PCIE

- BE/FE 테스트 환경

- Nginx를 이용한 port 분리
- frontend, backend, model 동일 머신
- github action 을 통한 빌드 자동화

- FE 주요 프레임워크

- Node v14.16.0
- Vue.js v2.6.14

- BE 주요 프레임워크

- python 3.8.10
- fastapi 0.78.0
- torch 1.11.0
- transformers 4.18.0

협업 툴

코드 관리 : github,huggingface hub

커뮤니케이션 : slack

아카이브: 노션 & 드라이브

The screenshot shows a GitHub repository page for 'final-project-level3-nlp-03'. The repository has 14 branches and 0 tags. The main branch is active. The page displays a list of recent commits, including:

- xuio-0528 Merge pull request #28 from boostcampaitech3/eunki/final ... b3a6e82 11 hours ago 127 commits
- backend feat: sim_score 보정 yesterday
- data_collection feat: pipeline 이용+직접 모델 선언해서 추론파일 3 days ago
- finetunning fix yesterday
- frontend/autograder add : done before graph 6 days ago
- keyword_checker feat: key_word가 주어진 상태일 때 그 key_word로만 비교하는 base checker 9 days ago
- modeling FastText 업로드 실험 12 hours ago
- prototype test: 데이터 읽기 노트북 3 days ago
- README.md first commit 17 days ago

The repository also includes a README.md file and a project description titled 'final-project-level3-nlp-03'.



프로젝트 : 곰파다

↳ 1개의 백링크



💡 다음 회의까지 준비사항

- 화: 린캠퍼스 1~4번까지 모두 작성
- 금: 서베이 시트 내용 정리(UI 저, 상렬, 세연 / 실험 계획 응기 / 데이터 진행상황 수정)
 - * 서베이 : 고객은 어느 방향, 어떤 기능이 추가되는지, 어떻게 작동했으면 좋겠는지 - 수치적 확인 목적
 - * 5월 3째주 정도에 마무리해서 공유하는 것으류~

💡 (프로젝트 시작되고나서) 금주 목표 : (언제까지) (어떤 것을) 완료

😎 네비게이션

- ▶ [프로젝트 사전 조사 내역 블락링크](#)
- ▶ [Timeline : 코어로 준비할 수 있는 기간 5월 3째, 4째, 5째주~6월 첫째주](#)

프로젝트 메인



데이터



개발



슬라이드 자료(미팅, 발표)



프로젝트 흐름



이활석 마스터님 피드백



로고

프로젝트 유ти



비즈니스 서치



컨택트 리스트



파일럿 학습용 데이터 리스트



프로젝트 서비스 구상도(기안)



Helpful Materials



비즈니스 유ти / 린캠퍼스

Timeline

	3월	4월	5월 1W	5월 2W	5월 3W	5월 4W	6월 1W
아이디에이션 & 사전 연구 조사	- 아이디어 제안 - 채점 관련 서비스 조사						
연구조사 심화			- 채점 기준 조사, 기 연구 조사 - 하민수 & 최성철 교수님 컨텍				
기능 정의				- UI설계 주요 기능 정의 - API 문서 정의, 개발 범위 분배			
프론트앤드 구현					- 주요 기능 mockup 테스트		
백엔드 구현					- API 문서 구체화, 백엔드 로직 구현 및 기능 추가 - FE & BE 연결 및 통신 확인		
벤치마크 기준 모델 파일럿 테스트				- BERT 및 오픈 데이터 조사 - 회귀, 분류로 파일럿 훈련			
Validation & 커스텀 데이터 구축				- Validation 용 pilot tagging 및 직접 구축 - 훈련용 데이터 크롤링 및 수작업			
키워드 모델 고도화						- word2vec 기반 키워드 체커	
유사도 모델 고도화						- SBERT 훈련 및 ablation study	
테스트						- 로직 구체화, 테스트	

파일럿 태깅 라벨링 기준

- 데이터 파악 및 Tagging 기준 통일을 위한 Pilot Tagging 진행

학생 답안	평가자1	평가자2	평가자3	평가자4	평가자5
점점 <u>공기</u> 의 온도가 낮아지고 <u>기압</u> 이 낮아지기 때문일 것 같다.	2	2	2	3	2
<u>기압</u> 이 높아져서 <u>공기</u> 의 압력때문에 <u>빵빵</u> 해지게 될 것일 것 같다.	1	0	0	0	0
고도가 높아지면 <u>기압</u> 이 낮아져서 과자 봉지의 내부 <u>공기</u> 가 <u>팽창</u> 한다	3	3	3	3	3

*Label 기준

3 = 답변에 포함된 키워드에 대한 모든 논리가 모범답안과 유사하거나 일치함

2 = 키워드에 대한 논리가 틀리다고 할 순 없지만, 맞았다고 하기도 애매함

1 = 키워드에 대한 논리가 틀린 개 1개 이상이고, 맞는 부분도 1개 이상 존재

0 = 키워드에 대한 논리가 모두 틀림

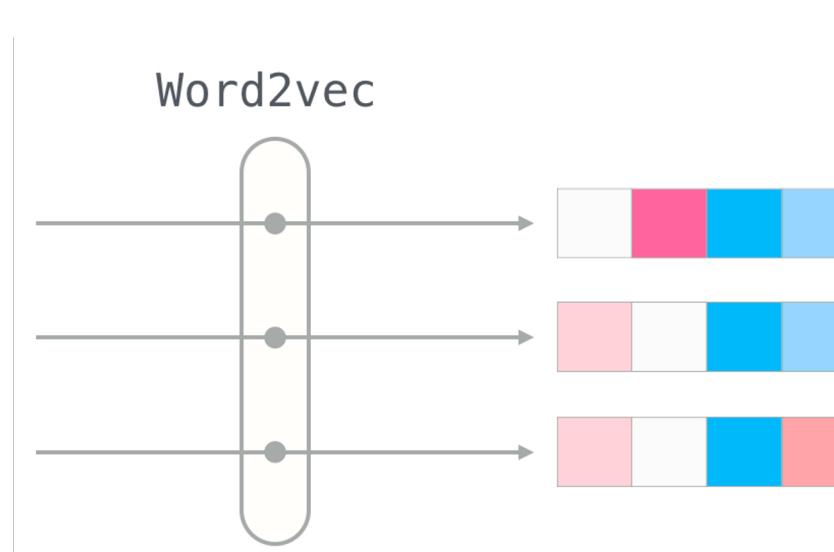
키워드 유사도 모델 진행 과정

- ① 교사가 제시한 키워드 벡터화
- ② 각 문장에서 띄어쓰기로 단어들을 구분
- ③ 형태소 분류기로 단어를 분리한 후 Noun, Verb, Adjective, Adverb들만 벡터화 및 유사도 검출
- ④ Threshold 이상일 경우 교사가 제시한 키워드와 관련이 있는 키워드로 분류
- ⑤ 점수에 포함시킨다.

키워드 :

가격, 품질, 다양성

값
품질
다양성



키워드 유사도 모델 진행 과정

- ① 교사가 제시한 키워드 벡터화
- ② 각 문장에서 띄어쓰기로 단어들을 구분
- ③ 형태소 분류기로 단어를 분리한 후 Noun, Verb, Adjective, Adverb들만 벡터화 및 유사도 검출
- ④ Threshold 이상일 경우 교사가 제시한 키워드와 관련이 있는 키워드로 분류
- ⑤ 점수에 포함시킨다.

키워드 :

가격, 품질, 다양성

학생 답안 :

빵을 더 낮은 값에 살 수 있고.. -----

split

빵을 더 낮은 값에 살 수 있고..

✓ ✓ ✓ ✓ ✓ ✓ ✓

키워드 유사도 모델 진행 과정

- ① 교사가 제시한 키워드 벡터화
- ② 각 문장에서 띄어쓰기로 단어들을 구분
- ③ **형태소 분류기로 단어를 분리한 후 Noun, Verb, Adjective, Adverb들만 벡터화 및 유사도 검출**
- ④ Threshold 이상일 경우 교사가 제시한 키워드와 관련이 있는 키워드로 분류
- ⑤ 점수에 포함시킨다.

키워드 :

가격, 품질, 다양성

학생 답안 :

빵을 더 낮은 값에 살 수 있고..



키워드 유사도 모델 진행 과정

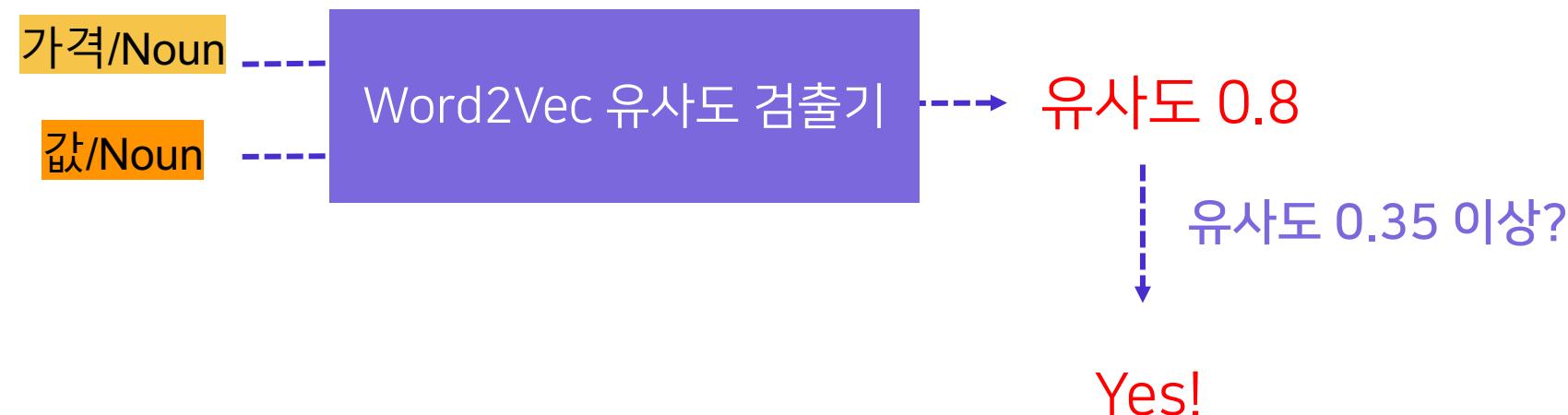
- ① 교사가 제시한 키워드 벡터화
- ② 각 문장에서 띄어쓰기로 단어들을 구분
- ③ 형태소 분류기로 단어를 분리한 후 Noun, Verb, Adjective, Adverb들만 벡터화 및 유사도 검출
- ④ Threshold 이상일 경우 교사가 제시한 키워드와 관련이 있는 키워드로 분류 → 점수 포함

키워드 :

가격, 품질, 다양성

학생 답안 :

빵을 더 낮은 값에 살 수 있고..



왜 직접 구현이 아닌 fastText 선택?

- 직접 구현한 Word2Vec의 limitation

- ① Corpus 단어 개수 부족
- ② 예측하지 못한 답이 나타났을 때 OOV 현상
- ③ 기존 학생 답안 위주의 학습으로 새로운 문제에 대한 유사도 분석 능력이 높지 않음



fastText

The word "fastText" is written in a bold, sans-serif font. The letters are colored red and blue. The "f" and "a" are red, while the "s", "t", "T", and "e" are blue.

- fastText의 장점

- ① n-gram 기법으로 OOV가 뜨지 않음
- ② 한국어 데이터 기반으로 매우 큰 word2vec이 이미 학습됨

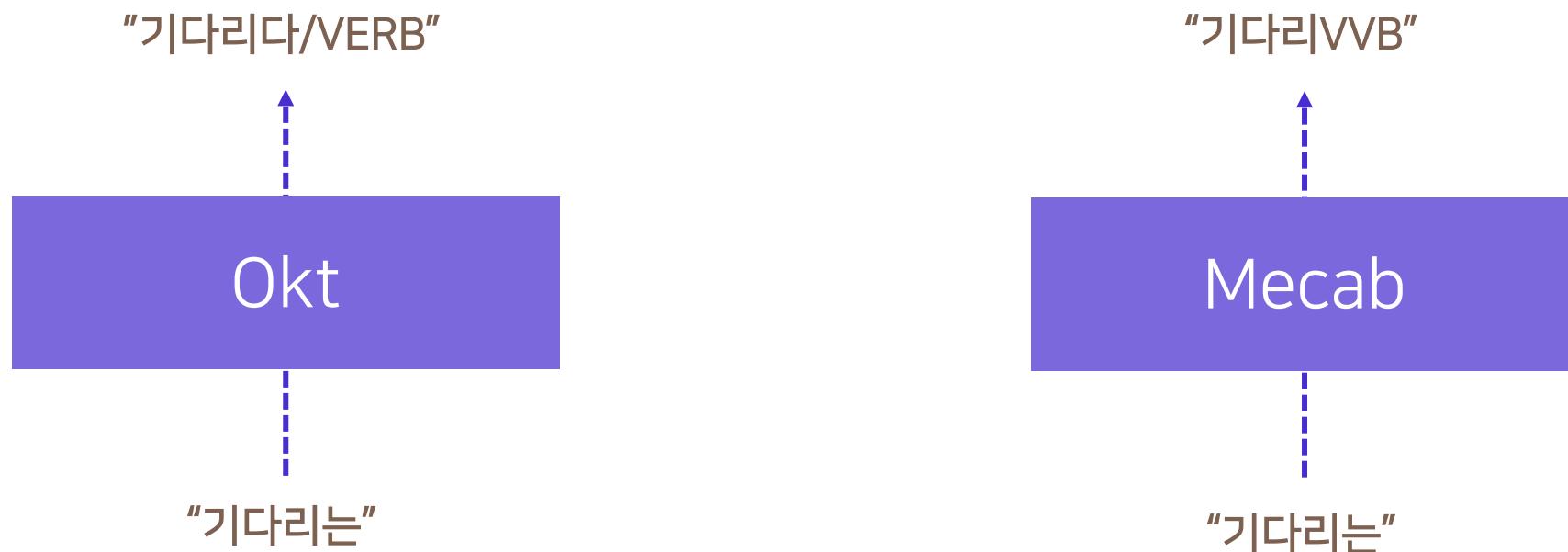
왜 문맥 단어 그대로가 아닌 Tokenizer를 사용했는가?

- 키워드 유사 모델에서는 어미, 조사를 제외하도록 구현하려고 했기 때문
Tokenizer를 통해 어미와 조사를 분리하지 않으면 단어의 의미가 달라지는 경우 존재



왜 Tokenizer 중에서 Okt를 선택?

- 유사 단어 추출 시 Noun, Verb와 같은 간단한 품사 정보만 필요했기에!
- 또한 Tokenization을 통해 lemmatization 형태로 빠르게 복원해주기 때문에



키워드 유사도 모델의 한계점과 개선방안

- 교사가 제시한 키워드에 대해 동일 혹은 유사 키워드의 존재 여부만 검사할 수 있다.

즉, 키워드가 문맥에서 어떻게 사용이 되었는가는 평가해내지 못함

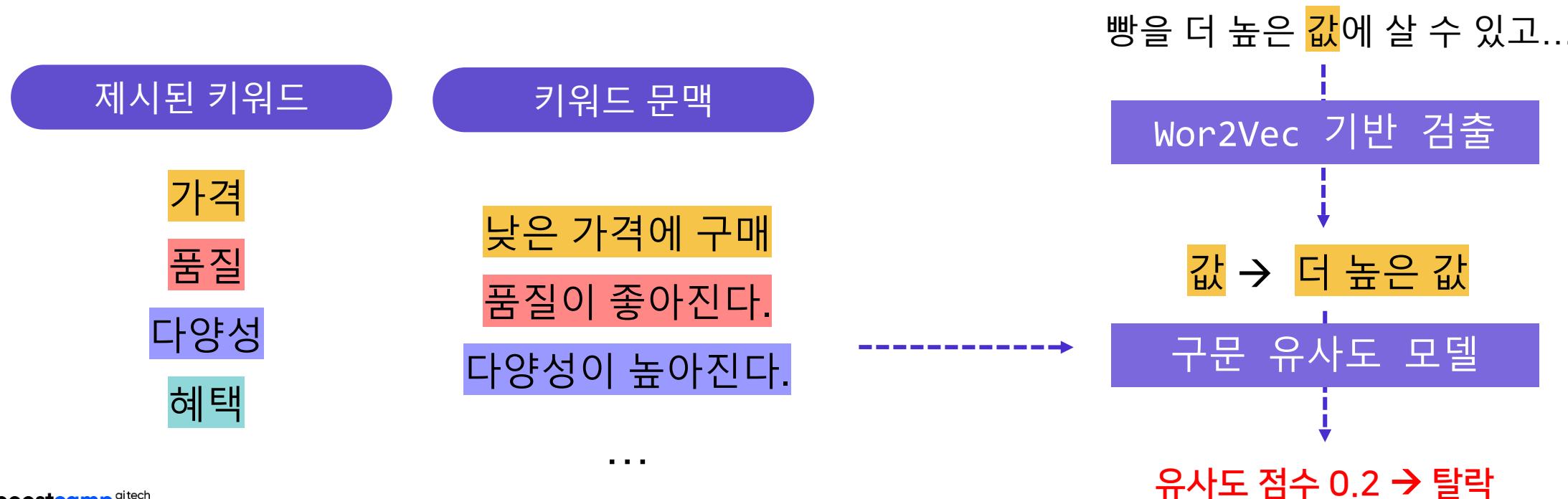


키워드 유사도 모델의 한계점과 개선방안

- 동일 혹은 유사 키워드가 검출되었을 경우 구문 형식의 유사도 검증 방식

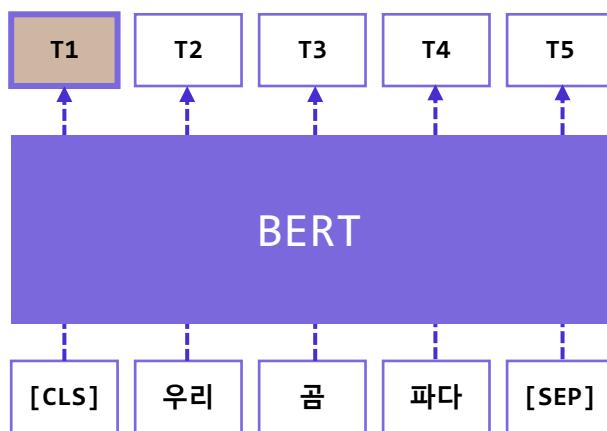
ex1) n-gram 방식을 차용해 구문과 모범답안의 키워드 문맥을 비교

ex2) 주체, 객체, 서술어 추출 방식 → 구문 검출

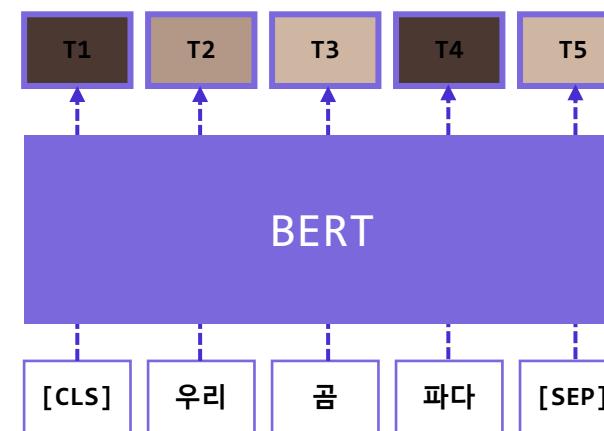


문맥 유사도를 어떻게 효과적으로 파악할까?

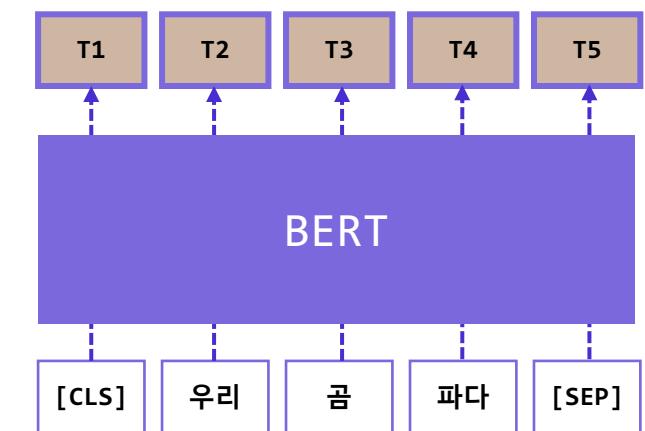
- BERT로부터 문장 벡터를 얻을 수 있는 방법



A. [CLS] 토큰



B. 토큰 Max pooling

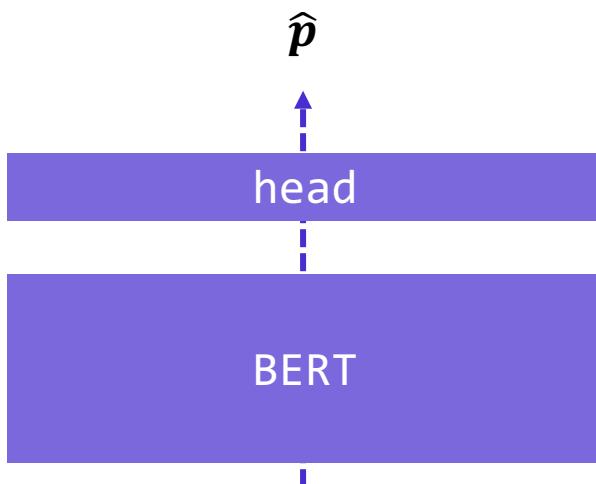


C. 토큰 Mean pooling

문맥 유사도를 어떻게 효과적으로 파악할까?

- 기본적인 문장 유사도 판단 학습 **limitation**

- ① 문장 길이 제약
- ② 모든 pair에 대해 계산 필요
- ③ 두 문장을 합쳐서 학습하는 것이 문장의 의미를 잘 파악하고 분류하고 있는 걸까?



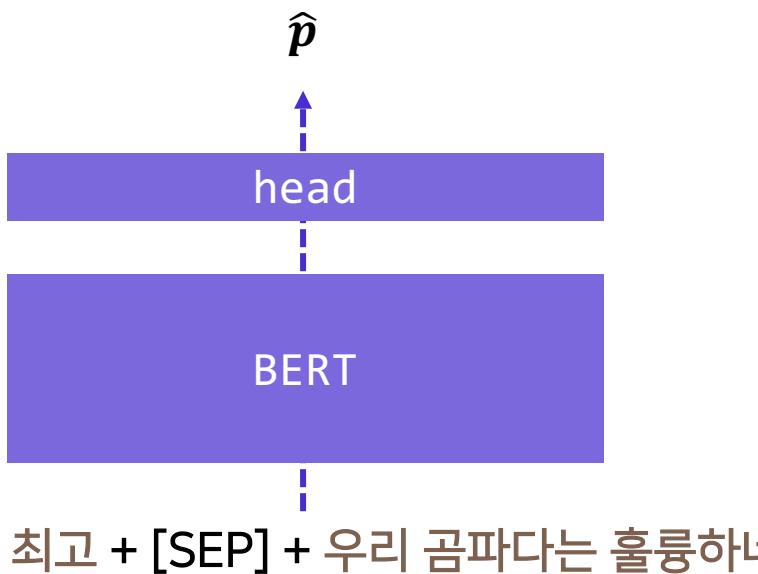
우리 곰파다 최고 + [SEP] + 우리 곰파다는 훌륭하네

- ① 일반적인 BERT의 token max length 512. 두 개 문장 합쳐서 512 되어야 함으로 문장 길이 제약

문맥 유사도를 어떻게 효과적으로 파악할까?

- 기본적인 문장 유사도 판단 학습 **limitation**

- ① 문장 길이 제약
- ② 모든 pair에 대해 계산 필요
- ③ 두 문장을 합쳐서 학습하는 것이 문장의 의미를 잘 파악하고 분류하고 있는 걸까?



① 선생님 답변 N개, 학생 답변 M개 \rightarrow NM번 계산 필요

문맥 모델 추가 예시 : 짧은 모범답안 1

모범 답안 : 지구는 주변을 돈다.

학생 답안	Cosine similarity	Threshold 0.6
지구는 주변을 돌지 않는다. 부정 표현 Catch!	0.58	0
주변을 지구는 돈다. 어순이 달라져도 Catch!	0.97	1
지구는 멈춰있다. 반의어 Catch!	0.54	0
멈춰있지 않고 중심을 기준으로 계속 움직인다.	0.47	0
지구는 멈춰있지 않고 중심을 기준으로 계속 움직인다.	0.76	1

같은 표현이지만 겹치는 단어가 없을 경우 놓침! '지구'라는 단어 포함 여부에 따른 성능차이 큼

문맥 모델 추가 예시 : 짧은 모범답안 2

모범 답안 : 콩팥은 재흡수, 분비 과정을 거쳐 오줌을 통해 노폐물을 몸밖으로 내보낸다.

학생 답안	Cosine similarity	Threshold 0.6
콩팥은 재흡수를 거쳐 오줌을 통해 노폐물을 몸밖으로 내보낸다.	0.99	1
콩팥은 재흡수를 거쳐 오줌을 통해 노폐물을 체외로 보낸다.	0.97	1
신장은 재흡수를 통해 오줌을 통해 노폐물을 체외로 보낸다.	0.91	1
소변을 통해 체외로 노폐물이 빠지는 과정은 신장에서 수행된다.	0.85	1
몸에 필요한 것들은 재흡수가 되고 불필요한 것은 신장에서 내보낸다.	0.71	1
노폐물들이 오줌을 통해서 체외로 내보내진다.	0.82	1
콩팥은 노폐물 배출을 담당하지 않는다. 반의어에 대한 파악을 어려워함	0.46	0
콩팥은 재흡수, 분비 과정을 거치지 않고 오줌을 통해 노폐물을 몸밖으로 내보낸다.	0.76	1
콩팥은 재흡수, 분비 과정을 거쳐 오줌을 통해 노폐물을 몸 안으로 재흡수한다.	0.97	1
콩팥은 재흡수, 분비 과정을 거쳐 오줌을 통해 중요한 성분을 몸 밖으로 내보낸다.	0.96	1
소변은 간과 연관되어 있고 노폐물들은 크게 영향을 주지 않는다.	0.40	0
모르겠어요.	0.00	0

문맥 모델 추가 예시 : 긴 모범답안

모범 답안 : 콩팥은 재흡수, 분비 과정을 거쳐 오줌을 통해 노폐물을 몸밖으로 내보낸다. 즉 항상성 유지를 위해 호르몬을 생산, 분비하는 내분비 기능을 가지고 있다.

학생 답안	Cosine similarity	Threshold 0.6
콩팥은 재흡수를 거쳐 오줌을 통해 노폐물을 몸밖으로 내보낸다.	0.85	1
콩팥은 재흡수를 거쳐 오줌을 통해 노폐물을 체외로 보낸다.	0.83	1
신장은 재흡수를 통해 오줌을 통해 노폐물을 체외로 보낸다.	0.79	1
소변을 통해 체외로 노폐물이 빠지는 과정은 신장에서 수행된다.	0.81	1
몸에 필요한 것들은 재흡수가 되고 불필요한 것은 신장에서 내보낸다.	0.69	1
노폐물들이 오줌을 통해서 체외로 내보내진다.	0.72	1
콩팥은 노폐물 배출을 담당하지 않는다.	0.37	0
소변은 간과 연관되어 있고 노폐물들은 크게 영향을 주지 않는다.	0.36	0
모르겠어요.	0.00	0

비교 기준이 되는 모범 답안이 더 많은 정보를 포함하고 길어질 수록 학생 답안 비교 점수가 떨어짐

성능 분석 : Confusion Matrix

- False Negative가 크게 감소 : 하지만 채점 입장에서 False Positive 감소가 더 중요하다..! 모델 & 데이터 구축에 Further works 필요!(TT)

		Predicted	
		Pos	Neg
Ground truth	Pos	36	21
	Neg	21	32

BERT 모델 + 회귀

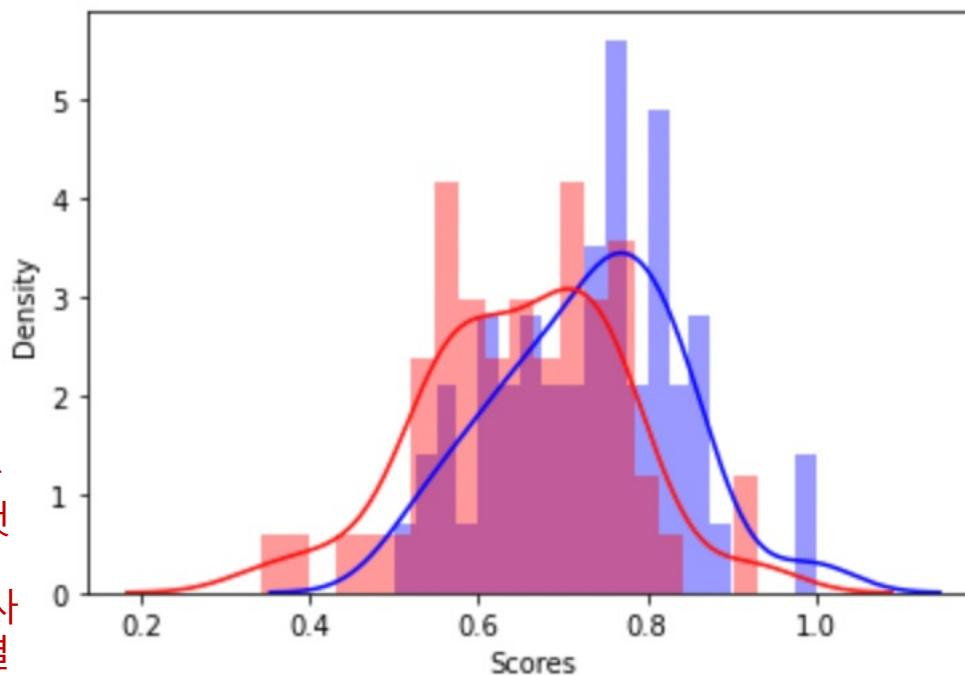
		Predicted	
		Pos	Neg
Ground truth	Pos	50	7
	Neg	29	24

SBERT 모델 + Cosine Loss

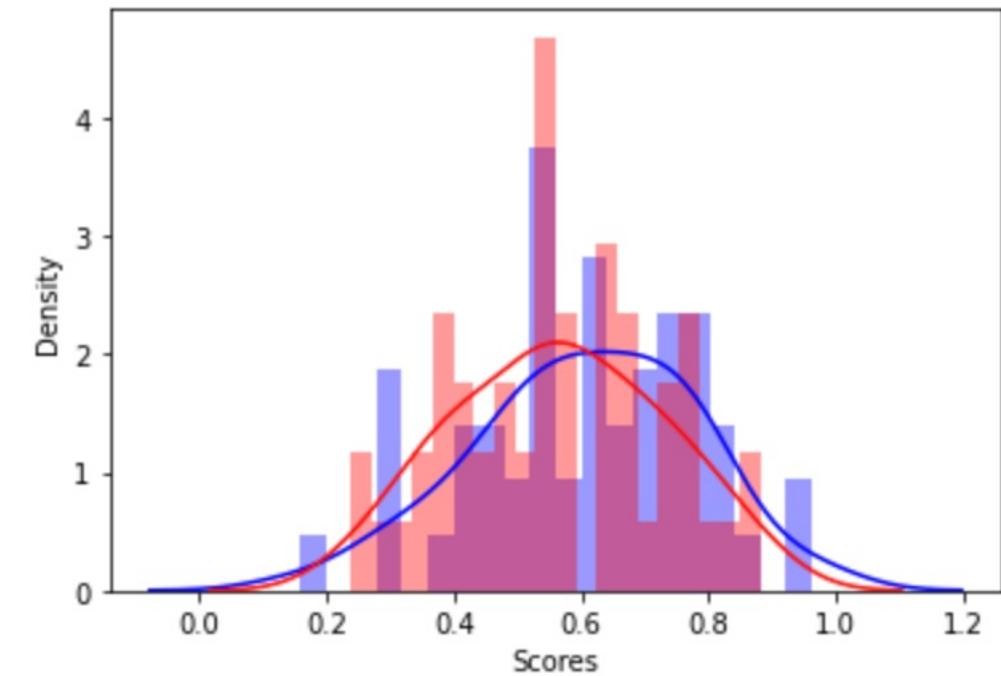
성능 분석 : 예측 분포

- SBERT의 FN이 감소한 이유는 회귀에 의한 결과값보다 유사도 점수의 분포가 좀더 높은 값을 가지기 때문

BERT
회귀 결과값
SBERT
유사도 값
* 회귀 점수와 유사도값을 비교하는 것은 무리가 있을 수 있으나, 최종값을 사용하는 관점에서 결과 분석 위해 사용



실제 동일한 pair에 대한 예측



실제 다른 pair에 대한 예측

데이터 : 문맥 유사도 채점 : 오픈 데이터 특징

Train

- 오픈 데이터

	KorSTS[1]	paraKQC[2]	Kor-sentence[3]	KLUE STS[4]
라벨	0~5 값	0 또는 1	0 또는 1	1. 0 또는 1 2. 0~5 값
특징	짧은 문장. 외국 STS-B 번역. 뉴스, 표현 설명 내용	짧은 문장. 질문중심	짧은 문장. 지식인 질문 포함. 인터넷 용어 다수	짧은 문장. Airbnb, Policy, paraKQC 포함
데이터 개수	5,749	15,170	61,220	11,668

- Rule based 제작 데이터 : 유의어, 반의어 데이터

- 라벨 : 0 또는 1
- 특징 : 짧은 문장 / 유의어, 반의어 Pair를 만들어 데이터 제작
- 데이터 개수 : 14,390

커스텀 데이터셋 제작 과정 : 유의어 및 단어 설명

구축 이유 : 학생 답변 중 유의어 파악을 잘 시키면 좋은 임베딩이 되지 않을까?



기존 자동 채점 연구 데이터[1]에서 발생하는 명사 추출(총 3,963개)

어학사전 Q

국어사전 단어 1-5 / 1,910건

나무¹ ▶ ★★★

[명사]

1. 줄기나 가지가 목질로 된 여러해살이 식물.
2. 집을 짓거나 가구, 그릇 따위를 만들 때 재료로 사용하는 재목.
3. 땔감이 되는 나무.
〔유의어〕 땔감, 땔나무, 목본

'명사'의 경우 뜻 풀이로 크롤링
단어집 구성

word	type
새풀	NNG
합병증	NNG
머름	NNG
그래픽	NNG
자기력선	NNG
분파	NNG
반도라	NNG
벤페	NNG

새로 갓 나온 파일이나 생선 따위를 이르는 말

어떤 질병에 곁들여 일어나는 다른 질병

바람을 막거나 모양을 내기 위하여 미닫이 문지방 …

그림이나 도형, 사진 등 다양한 시각적 형상이나 작…

자기장의 크기와 방향을 나타내는 선

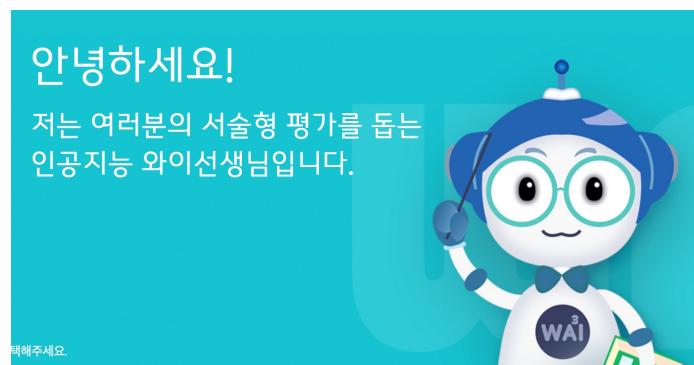
일을 잡쳐서 실패함.

기타와 비슷하며 낮은 소리를 내는, 르네상스 시대…

거죽에 곱고 짧은 털이 촘촘히 돋게 짠 비단.

커스텀 데이터셋 제작 과정 : 반의어

구축 이유 : 학생 답변 중 중요한 동사가 정반대의 의미를 갖는 경우가 다수!!



기존 자동 채점 연구 데이터[1]에서
발생하는 용언(577), 형용사(112) 추출

어학사전

국어사전 단어 1-5 / 1,910건

나무¹ ★★★

(명사)

- 줄기나 가지가 목질로 된 여러해살이 식물.
- 집을 짓거나 가구, 그릇 따위를 만들 때 재료로 사용하는 재목.
- 땔감이 되는 나무.

(유의어) 댤감, 댤나무, 목본

반의어 부재

전문가 감수 정보 [?] 참여자 제안 정보 [?]

내려-가다 편집하기 편집 금지 요청

발음 [내려가다]
활용 내려가 [내려가], 내려가니 [내려가니]
품사/문형 「동사」 […에][…으로]

- 「001」 높은 곳에서 낮은 곳으로 또는 위에서 아래로 가다.
▶ 아래층에 내려가다.
▶ 아버지 대신에 아들이 공구를 찾으러 지하실에 내려갔다.
▶ 그는 강터를 빠져나와 신작로를 따라 아래쪽으로 내려갔다.
▶ 이윽고 일어선 그는 엘리베이터를 타고 곧장 지하로 내려갔다.《박완서, 오만과 용상》

해당 단어 부재

관련 어휘

- 반대말 올라-기다
- 지역어 (방언) 나래-기다(강원), 나려-기다(경기), 니려-기다(경남), 느려-기다(경북), 니리-기다(경북), 내려-기다(경상, 전남), 내레-기다(경상, 전남, 평안), 나리-기다(전남), 내라-기다(전남), 내리-기다(전남), 네리-기다(전남), 니라-기다(전남)
- 옛말 내려-기다

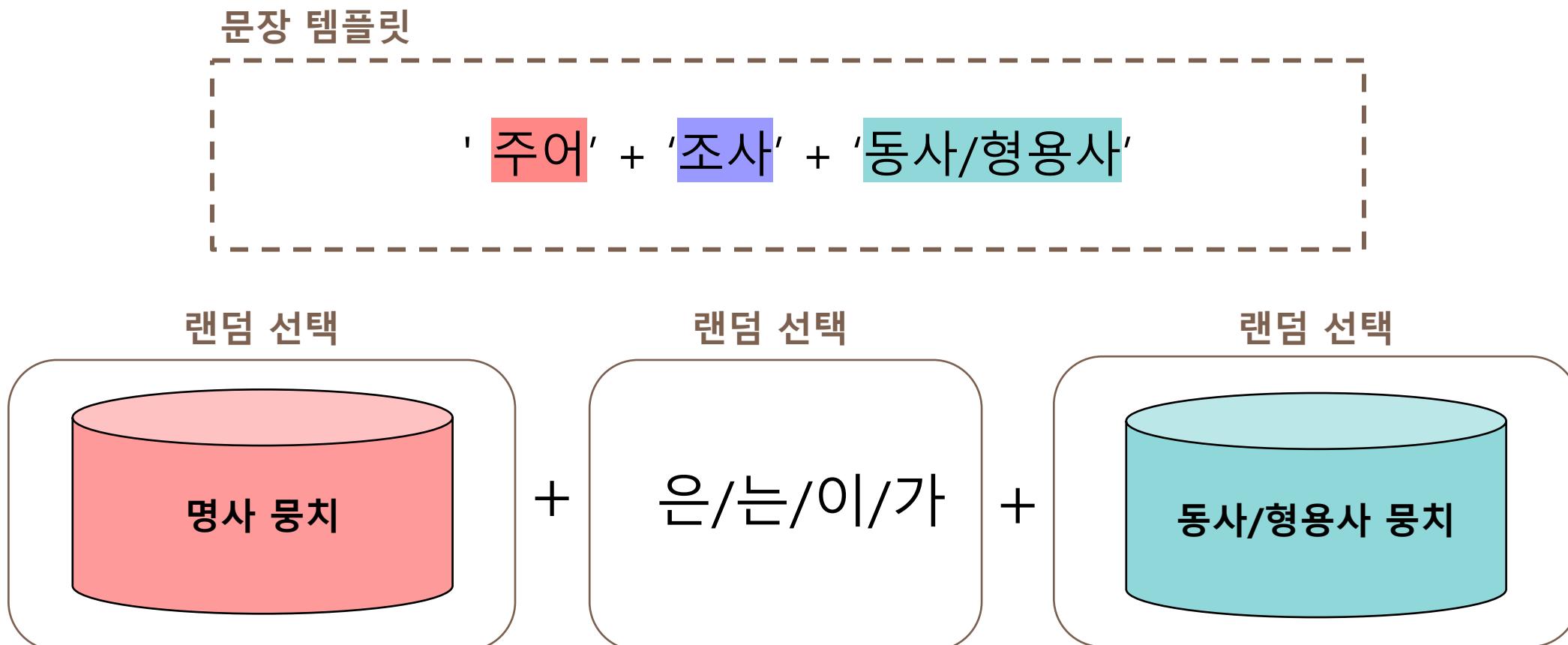
커스텀 데이터셋 제작 과정 : 반의어

- 번역 과정에는 Papago Translation API, 반의어 검색에는 시소러스 영어사전 사용
- 복잡한 우회 과정 :** 한국어 --> 영어 --> 영어 반의어 --> 한국어 반의어



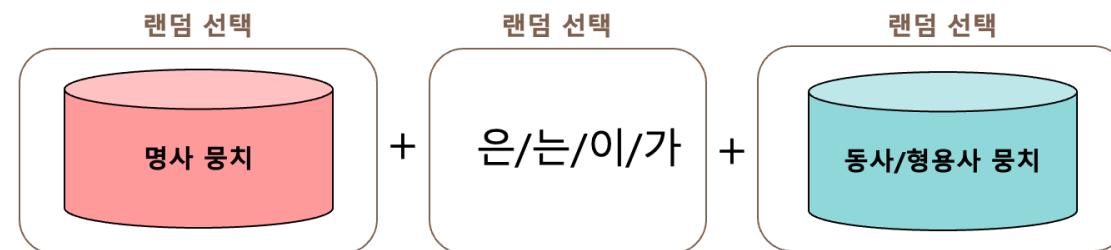
커스텀 데이터셋 제작 과정 : 데이터 생성

구축 이유 : 모아진 유의어, 의미 풀이, 반의어로 데이터를 생성해서 학습하면 더 효과적인 negative sample을 만들 수 있지 않을까?



커스텀 데이터셋 제작 과정 : 데이터 생성

구축 이유 : 모아진 유의어, 의미 풀이, 반의어로 데이터를 생성해서 학습하면 더 효과적인 negative sample을 만들 수 있지 않을까?



생성된 원래 문장 : **하늘은 푸르다**

Positive Samples

지평선이나 수평선 위로 보이는 무한대의 넓은 공간은 푸르다

지평선이나 수평선 위로 보이는 무한대의 넓은 공간은 생활에서 충분한 만족과 기쁨을 느끼어 흐뭇하다.

Negative Samples

하늘은 안 푸르다

하늘은 누렇다

지평선이나 수평선 위로 보이는 무한대의 넓은 공간은 안 푸르다

성능 Ablation : 데이터셋 종류에 따른 성능

예측값이 threshold 이상일 경우 동일한 문장, 미만일 경우 다른 문장으로 처리 후 binary accuracy 책정

데이터셋				Threshold			
korSTS	paraKQC	생성데이터	kor-sentence	0.6	0.7	0.8	0.9
✓				0.62	0.59	0.48	0.50
✓	✓			0.59	0.60	0.55	0.53
✓		✓		0.55	0.50	0.48	0.43
✓			✓	0.53	0.47	0.45	0.48
	✓			0.61	0.61	0.63	0.59
		✓		0.52	0.52	0.52	0.52
			✓	0.58	0.58	0.57	0.56

korSTS 외에 데
이터셋은 binary
임으로 실수정
보를 주는
korSTS가 높은
성능을 준 것으
로 예측

모델 : klue/bert-base

BS: 32, lr: 2e-5 (허깅페이스에 보고된 일반적 STS 학습에 사용한 configuration을 따름[1])

성능 Ablation : 데이터셋 종류에 따른 성능

예측값이 threshold 이상일 경우 동일한 문장, 미만일 경우 다른 문장으로 처리 후 binary accuracy 책정

데이터셋					Threshold			
korNLI	korSTS	paraKQC	생성 데이터	koSTS 반의어 치환	0.6	0.7	0.8	0.9
pretrained					0.64	0.59	0.62	0.52
	✓				0.61	0.63	0.65	0.52
✓	✓				0.66	0.61	0.59	0.53
✓		✓			0.60	0.54	0.55	0.55
✓			✓		0.60	0.56	0.55	0.53
✓	✓			✓	0.52	0.52	0.52	0.52

모델 : stsbert-xlm-r-multilingual

BS: 32, lr: 2e-5 (허깅페이스에 보고된 일반적 STS 학습에 사용한 configuration을 따름[2])

성능 Ablation : 데이터셋 종류에 따른 성능

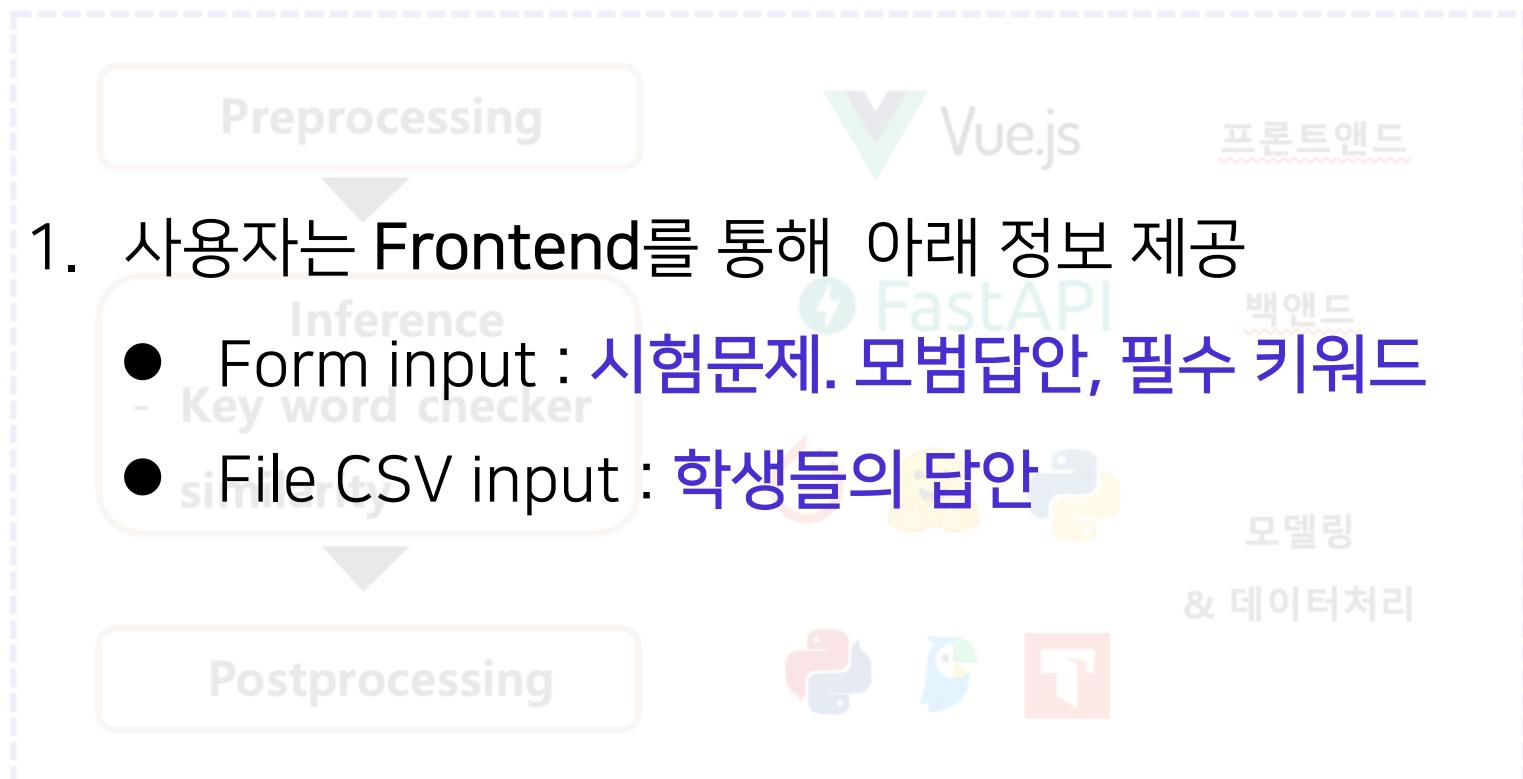
예측값이 threshold 이상일 경우 동일한 문장, 미만일 경우 다른 문장으로 처리 후 binary accuracy 책정

데이터셋						Threshold			
korNLI	korSTS	klueSTS	paraKQC	생성 데이터	koSTS 반의어 치환	0.6	0.7	0.8	0.9
	✓					0.60	0.59	0.54	0.47
✓	✓					0.67	0.65	0.58	0.48
✓		✓				0.63	0.58	0.55	0.50
✓	✓					0.53	0.61	0.61	0.56
✓			✓			0.64	0.59	0.52	0.48
✓				✓		0.65	0.64	0.59	0.57
✓	✓				✓	0.52	0.52	0.52	0.52

모델 : klue/bert-base

BS: 32, lr: 2e-5 (SKT sentence-transformer STS 학습에 사용한 configuration을 따름[1])

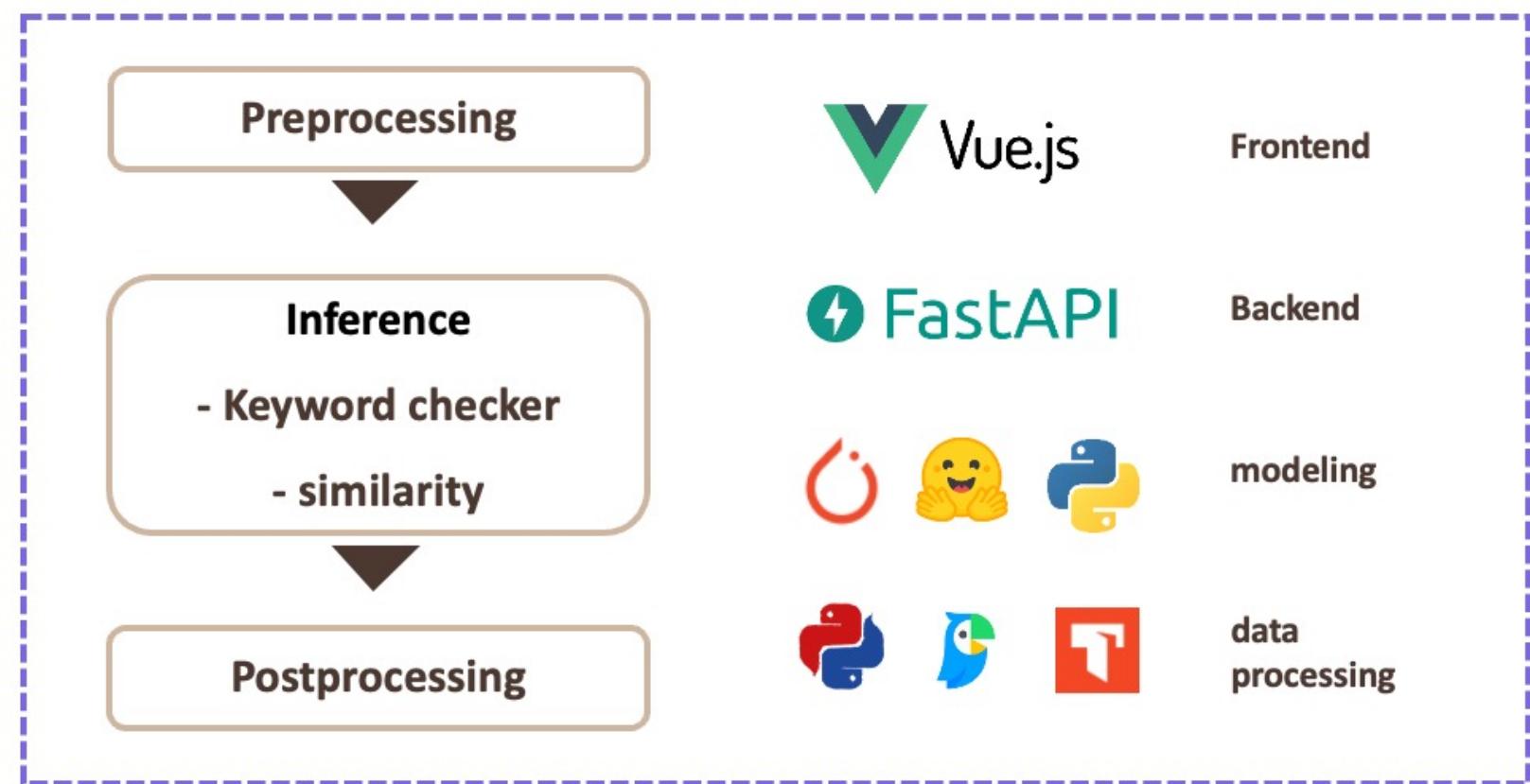
유저 시나리오



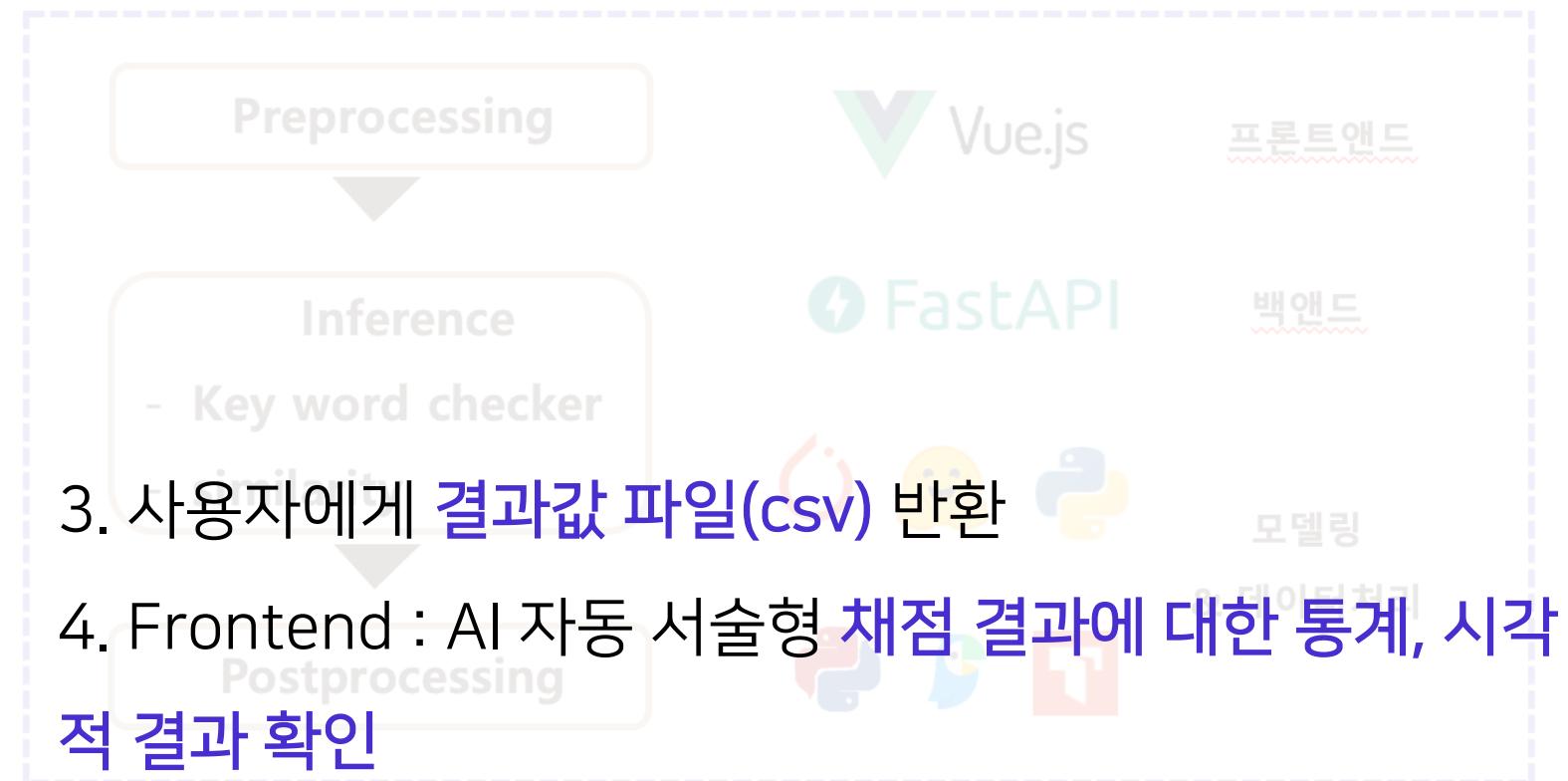
유저 시나리오



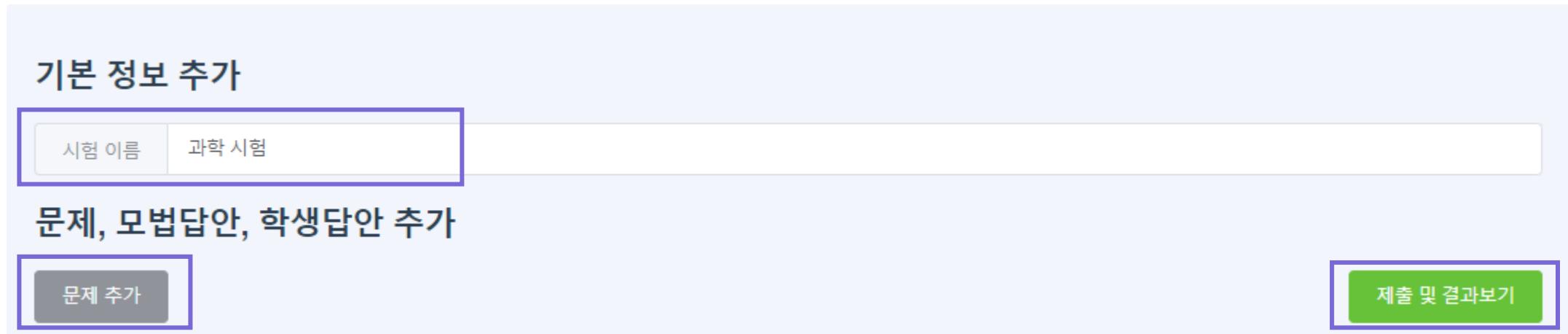
2. FastAPI에서 정보를 받아 전처리, Text-similarity& 키워드 모델로 결과값 산출



유저 시나리오



프로덕트 기능 - 기본정보 추가



- ① **시험 이름** 추가 가능
- ② **문제 추가** 버튼을 통해 채점할 문제 추가 가능
- ③ **제출 및 결과보기** 버튼을 통해 채점 결과 확인

프로덕트 기능 - 문제 추가 디테일

문제 1 ×

문제와 모범 답안

문제 뜨거운 물이 들어있는 냄비에 나무국자와 금속 국자를 넣었는데, 나무국자는 뜨거워지지않고 금속국자는 뜨거워지는 이유는 무엇일까요?

모범답안 나무 재질보다 금속 재질이 열 전달이 더 잘 되기 때문이다.

필수 키워드

열 × 전달 × 재질 × 나무 × 금속 × + New Tag

파일 업로드

Drop file here or [click to upload](#)

only csv files allowed

q1.csv

① 문제와 모범답안, 필수 키워드 입력

② 학생 ID와 학생 예시 답안이 포함된 csv파일 업로드

demo_data > q1.csv

```
1 student_id,answer
2 0,나무보다 금속이 열 전달을 더 잘하기 때문에
3 ,금속이 나무보다 열을 더 잘 흡수를 잘하기 때문이다.
4 12,금속이 나무보다 열전도성이 좋기 때문이다.
5 18,나무 국자보다 금속 국자가 열 전도성이 좋기 때문이다.
6 19,금속이 나무보다 열을 더 잘 이동시키기 때문
7 20,나무보다는 금속이 더 잘 뜨거워 지기 때문이다.
```

csv파일 예시

프로덕트 기능 - 결과 확인

결과 확인		
Index	Student ID	Score
1	0	86 / 100
2	3	76 / 100
3	12	72 / 100
4	18	65 / 100
5	19	73 / 100

- ① 각 학생 별 서술형 AI 채점 점수를 확인 가능
- ② 자세히 보기 : 어떤 방식의 채점이 적용됐는지 상세 사항 확인 가능

프로덕트 기능 - 결과 확인 상세

채점 결과 X

문제번호	키워드	세부 점수	최종 점수
▼ 1	<p>열 전달 재질 나무 금속</p> <p>나무보다 금속이 열 전달을 더 잘하기 때문에</p>	<p>Keyword Score : 0.8 Similarity Score : 0.93</p>	<p>86 / 100</p>

문제 : 뜨거운 물이 들어있는 냄비에 나무국자와 금속 국자를 넣었는데, 나무국자는 뜨거워지지 않고 금속국자는 뜨거워지는 이유는 무엇일까요?
모범답안 : 나무 재질보다 금속 재질이 열 전달이 더 잘 되기 때문이다.

- ① 자세히 보기 : 채점 점수
- ② 키워드 : 학생의 답안에 포함된 모범 키워드
- ③ 세부점수 탭 : Similarity Score와 Keyword Score 각각 확인
- ④ 최종점수 탭 : 각 점수에 적절한 가중치를 곱해 최종 점수 산출